P6QLOPE1

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK
----------x

IN RE: OpenAI INC. COPYRIGHT INFRINGEMENT LITIGATION 25 MD 3143 (SHS)(OTW)

----------x

New York, N.Y.
June 26, 2025
10:00 a.m.

Before:

HON. SIDNEY H. STEIN,

District Judge

HON. ONA T. WANG,

Magistrate Judge

APPEARANCES

SUSMAN GODFREY LLP
Attorneys for Authors Guild and Alter Class Plaintiffs
BY: JUSTIN NELSON
CRAIG SMYSER
ROHIT NATH
ALEJANDRA SALINAS
CHARLOTTE LEPIC
JORDAN CONNORS
-and-
LIEFF CABRASER HEIMANN & BERNSTEIN LLP
BY: ANNA FREYMANN
-and-
BOIES SCHILLER FLEXNER LLP
BY: JOSHUA STEIN
-and-
JOSEPH SAVERI LAW FIRM
BY: CHRISTOPHER K. L. YOUNG

SUSMAN GODFREY LLP
Attorneys for Plaintiff The New York Times Company
BY: DAVIDA BROOK
IAN CROSBY
ALEX FRAWLEY
ADNAN MUTTALIB

APPEARANCES (continued)

ROTHWELL FIGG ERNST & MANBECK PC
Attorneys for Plaintiff Daily News LP and The New York Times Company
BY: STEVEN M. LIEBERMAN
JENNIFER MAISEL
KRISTEN LOGAN
JENNY COLGATE

LOEVY & LOEVY
Attorneys for CIR, The Intercept Media, and Raw Story
BY: MATTHEW TOPIC

KLARIS LAW
Attorneys for Ziff Davis Plaintiffs
BY: LANCE KOONCE

KEKER VAN NEST & PETERS LLP
Attorneys for Defendant OpenAI Inc., et al
BY: ROBERT VAN NEST
EDWARD BAYLEY
MICHELLE YBARRA
THOMAS GORMAN

LATHAM & WATKINS LLP
Attorneys for Defendants
BY: ANDREW GASS
-and-
ORRICK HERRINGTON & SUTCLIFFE LLP
BY: ANNETTE L. HURST
-and-
MORRISON & FOERSTER LLP
BY: JOSEPH C. GRATZ

Also Present: Guy Ruttenberg, Esq.

(Case called)

MR. NELSON: Good morning, your Honors. Justin Nelson representing the class plaintiffs. With me for the class plaintiffs, I have on the first row Craig Smyser from Susman Godfrey. On the second row, Rohit Nath from Susman Godfrey, Alejandra Salinas from Susman Godfrey. I also have in the back Charlotte Lepic from Susman Godfrey, Jordan Connors from Susman Godfrey, and Anna Freymann from Lieff Cabraser, and Joshua Michelangelo Stein from Boies Schiller & Flexner, and Christopher Young from the Joseph Saveri law firm.

JUDGE STEIN: Good morning to all of you.

MS. BROOK: Good morning, your Honor. Davida Brook on behalf of the News Plaintiffs as liaison counsel and The New York Times. And with me today with the News Plaintiffs is Ian Crosby of Susman Godfrey on behalf of The New York Times, Steven Lieberman of Rothwell Figg on behalf of The New York Times and The Daily News, Jennifer Maisel in the front well on behalf of Rothwell Figg and The New York Times and The Daily News, Kristen Logan of Rothwell Figg on behalf of The New York Times and The Daily News, Jenny Colgate from Rothwell Figg on behalf of The Daily News. Lance Koonce is in the gallery on behalf of Klaris Law and Ziff Davis. Guy Ruttenberg is here from Ruttenberg IP Law on behalf of Ziff Davis; Alex Frawley and Adnan Muttalib from Susman Godfrey on behalf of The New York Times, and Matt Topic from Loevy & Loevy on behalf of

the Center for Investigative Reporting and The Intercept.

THE COURT: Good morning to all of you.

MR. VAN NEST: Good morning, Judge Stein. Good morning, Judge Wang. Bob Van Nest of Keker, Van Nest & Peters here for OpenAI. I am here with Ed Bayley, Michelle Ybarra, Tom Gorman from our firm, Joe Gratz from Morrison & Foerster, Andy Gass from Latham & Watkins. And we have representatives from OpenAI here as well today, Renny Hwang, Susan Kim, and Molly Yates.

JUDGE STEIN: Thank you. Good morning.

MS. HURST: Good morning, your Honor. That leaves me, Annette Hurst on behalf of Microsoft. And that's it. Just me, your Honor.

JUDGE STEIN: That's wonderful. I assume only a small percentage of the names that have been called are billing time for today. Please be seated.

The purpose of this hearing is for the parties to go over the technology, you have been calling it a technology tutorial, so that Judge Wang and I have a better understanding of the large language models and exactly how the technology works. We did a little of it a few months ago, but as I said, when I set this conference, each time I hear it, I feel a little more comfortable with the technology and I learn a little more. The parties have asked for an hour each. I am not sure that's necessary, but you each have an hour. Let's

begin with the plaintiffs.

MR. NELSON: Good morning, your Honor. Justin Nelson from Susman Godfrey, may it please the Court, representing the class plaintiffs. I am joined by Jennifer Maisel from Rothwell Figg representing the news plaintiffs, and together we are going to present the plaintiffs' combined technology tutorial.

We will discuss the technical details in the context of the challenged use they relate to: First, the technical means by which Defendants made their initial copies of books and news articles; second, the copies of the copyrighted material made during the training process, and how the training process works; third, the large language model, or LLM, resulting from the training process, including the manner in which it stores copies of the copyrighted works; fourth, the LLM output, including reproductions of the copyrighted works used in the training process, in a process of so-called regurgitation; fifth and finally, the so-called retrieval augmented generation or RAG products from Defendants which may also access copy or display copyrighted works at issue in this case. The allegation is that each use constitutes copyright infringement, and that's what this case is about.

Over the next little bit, we will go through how large language models are created and how they work. We will also show the technical methods by which the defendants acquired and fed data to these models, how and why content from Class and

News Plaintiffs is critical to the functionality of these LLMs, and how the basic technology of LLMs manifests in the products and services from OpenAI at issue here.

On the screen is a book called *Split Second* by David Baldacci, who is one of the named plaintiffs. The allegations in this case are that OpenAI and Microsoft downloaded and built digital libraries made out of not just one book or one news article, but millions of books and millions of articles without permission. As to books, OpenAI and Microsoft purposely sought out and sourced these books from, among other sources, known pirated websites, including sites like Library Genesis. As to news content, the defendant sought out and sourced material by, among other things, purposefully ignoring terms of service, digital no-trespassing signs, and deploying software to circumvent paywalls --

JUDGE STEIN: Let me just stop you. I am going to ask questions periodically.

MR. NELSON: Of course.

JUDGE STEIN: Go back to the slide, prior slide. "OpenAI and Microsoft pirated copyrighted books without permission." You are saying they copied copyrighted books, right? Your take on it is it was a violation of copyright and, therefore, pirated? Is that the idea?

MR. NELSON: Correct, your Honor. I think, actually, there's no doubt that there is a copying here. I think the

defense that they will have is fair use, whether it's a proper fair use. But in terms of the --

JUDGE STEIN: I am with you so far. Go ahead.

MR. NELSON: In terms of the copying, yes, they went either by downloading -- I will get into this in a couple minutes -- torrenting books from these pirated websites --

JUDGE STEIN: All I was doing is I am trying to separate the facts from the advocacy.

MR. NELSON: I'm sorry.

JUDGE STEIN: So the "pirated" is the advocacy part here.

MR. NELSON: Well, I actually -- I don't think "piracy" is a term of advocacy. It is a term of art to talk about taking books or other copyrighted content without permission.

JUDGE STEIN: OK. OpenAI and Microsoft unlawfully copied copyrighted books without permission?

MR. NELSON: Well, your Honor, I think "unlawfully" would get -- that would be advocacy. I think "piracy" is a little bit different. "Unlawfully," that is their fair use defense, to be fair to them.

JUDGE STEIN: OK. Let's not get hung up on it. I'm sorry.

MR. NELSON: No. Thank you, your Honor.

So as to news content, as we were discussing, the

allegations are they did a number of things to deploy software, including circumventing paywalls to copy millions of articles. Throughout this presentation, we will refer to this as "AI," artificial intelligence, but "AI" is really a misnomer. It is HI, human imitation, as we will see. These models were designed to emulate the human expression they have copied and ingested through training.

The basic formula underpinning today's large language models is exactly what's on this slide. The quality of the model is based on, one, data, and two, compute. Because data and compute are so important and because the amount of data and compute used in training models is growing exponentially, OpenAI and Microsoft worked hand in glove to secure the two. Both OpenAI and Microsoft worked to source huge quantities of high-quality books and news data, and both OpenAI and Microsoft are spending many billions of dollars to develop, design, and implement the gigantic super computer clusters necessary to train and deploy the large language models at issue. It's not just OpenAI and Microsoft. There is no secret sauce. It is these two factors of production. All AI models are basically the outcome of these two factors, data and compute. At the heart of these models, their basic core --

JUDGE STEIN: Before you go on, what this slide, the basic formula slide, what I take from that is simply what the defendants do is they take data, they do something with it, and

the result is the large language model.

MR. NELSON: Yes. And as we will get to in a few minutes, your Honor, it is actually the more data you have and the more compute you have, the better your model is going to be. So it is a function of those two about how good your model is in your scaling laws that we will get into.

JUDGE STEIN: You will tell us what the "compute" part is as you go along.

MR. NELSON: Correct, your Honor. But just very simply, it is the power that allows these large language models to operate. It's why companies like NVIDIA who make these chips have now gone to the stratosphere in terms of value, because they sell the chips. You have data centers who are running servers, who are putting it on that have these calculations that are able to do in literal, you know -- in a second, they can do literally billions of calculations.

JUDGE STEIN: The more data -- you say the more data, the better data is based on the more data that comes in, the more able the large language model is to imitate human language?

MR. NELSON: That's exactly right, your Honor.

JUDGE STEIN: OK. Let's move on.

MR. NELSON: And so at the heart of these models is the data. And so this is what one of the founders of OpenAI, Ilya Sutskever, has said: "Data is the fossil fuel of AI. It

was, like, created somehow. And now we use it." Of course, we know it wasn't, like, created somehow. It was created by humans at enormous financial personal and physical cost. And the AI systems are trying to mimic it. In other words, the data used in developing these LLMs is human expression, which OpenAI and Microsoft used as the key factor of production. And as we will see, the quality of the expression in the writing is the key component of how good the model is. Books and news articles in particular are an incredibly valuable finite resource for AI companies. They represent the outcomes of human creative processes that cannot be faked.

The models we refer to as AI models are called that because they model, which is to say they approximate their training data. As one OpenAI researcher put it, the "it" in AI models is the data set. He says, "Models are truly approximating their data sets to an incredible degree."

JUDGE STEIN: The output of the model is trying to approximate what is contained in the data set?

MR. NELSON: That is exactly right, your Honor.

JUDGE STEIN: OK.

MR. NELSON: Model behavior is determined by your data set. Nothing else. Everything else is a means to an end in efficiently delivering compute to approximating that data set.

You will hear the defendants say that model -- the models are learning patterns and language, but that's not

right. What the model is doing is encoding human expression. The reason an LLM like those released by OpenAI and Microsoft can generate expressive content is because it's copied and encoded the expression from countless works like *Split Second* into its model of training data.

For example, we asked one of OpenAI's LLMs, "Write me the first chapter of a novel in the style of David Baldacci," and it produced a book called *The Orion Protocol*. The reason OpenAI's LLM can do that is because it's copied and encoded many, many books. The same is true for news articles. And without that high quality human data, the result is a low quality model. Garbage in, garbage out. How do OpenAI and Microsoft's products produce something that emulates human expression?

I will be explaining how these LLMs work from how the books and news material were acquired, to the importance of the quality expressive content to the model, and through the intricacies of how it encodes expression in order to emulate its training data. Ms. Maisel will then explain how OpenAI and Microsoft take these models and further refine them for usage in their products, how the models continue to output works seen during training, and how all of this actually works in the products at issue in these suits.

So let's talk about how an LLM is trained, including the technical methods OpenAI and Microsoft used to acquire the

training data and ready it for use in the training process.

There are three main steps to the LLM training process: Step 1 is acquiring the works and then isolating the expression into plain text. Step 2 is pre-training, where the data acquired in Step 1 is copied over and over again, millions, billions of times, to generate the base model. And Step 3 are the post-training steps, including fine-tuning, mid-training, and others that Ms. Maisel will address.

Let's start with data acquisition. OpenAI and Microsoft wanted to obtain huge libraries of high quality books and news articles as quickly and cheaply as possible. So where did they go? With copyrighted books, the answer for OpenAI and Microsoft was to seek out and download, more precisely, torrent millions of books from pirated sites. As the U.S. Government's 2022 Review of Notorious Markets for Counterfeiting And Piracy has recognized, online piracy is a term of art. It refers to "the practice of uploading, downloading, distributing, or streaming the digital copyrighted works over the internet without the permission of rights holders or other legal authorization." As that same report notes, torrents allow users to download files containing copyright-protected content over peer-to-peer networks and are a prominent delivery method used for downloading unlicensed content. Torrenting allows users to download millions of books at high speeds from places that are the equivalent of Napster for books. Those pirated

sites include Library Genesis, or LibGen, which the U.S. Government recognizes as a notorious market for counterfeiting and piracy, which has a pirate ship for its logo, and it is well-known within the tech community that LibGen is in blatant violation of copyright, and their domain name is blocked in some countries. In fact, LibGen and its related shadow libraries have been repeatedly shut down by courts in this district for piracy, and shadow libraries themselves have noted that AI has saved them from oblivion.

Here is a post from Notorious Pirate Library, Anna's Archive: "Not too long ago, 'shadow libraries' were dying. Then came AI." For both books and news content, OpenAI and Microsoft also obtained huge swaths of data via process colloquially known as scraping.

JUDGE STEIN: Before you go on to that, let me understand the torrents a little better. Torrents is not a specific website; it's a style?

MR. NELSON: It's an application, basically, your Honor, where it is distributed amongst hundreds, thousands, more, computers by which it's a peer-to-peer network so that they are stored on a variety of different computers, and you can upload or download from other computers, and it's a distributed network that allows for that to happen.

JUDGE STEIN: What's in it for Anna's Archive? Why would an entity be doing this?

MR. NELSON: Well, for a couple of reasons. First is this idea that all data should be free, all information should be free. There are ideological reasons for that, which you can't dismiss. The second is, some of them do receive some fee for, for example, speeding up that delivery process, and some of them are just --

JUDGE STEIN: You pay us, and we will give you 2 million books sooner than some other competing torrent? Is that the phrase?

MR. NELSON: Just -- you get in -- it's like a speed pass at Disney World. You get to go up a little quicker.

JUDGE STEIN: OK. Thank you.

MR. NELSON: So with respect to scraping, many sites are OK with scraping because they will show up in search results, but they often have strict restrictions on the display or use of the data obtained for that search engine purpose. These restrictions may be set forth in some combination of certain technical guidance embodied in a file called robots.txt, website terms of use which specify to what extent scraping is or is not permissible, and the use of paywalls.

Microsoft, for example, runs a huge scraping program as part of its Bing search engine. OpenAI has also run its own scraping programs and source scraped data from third parties. In the case of news content, while some was available without a paywall, large portions of the content was behind a paywall.

Rather than respecting traditional scraping processes and the restrictions embodied by paywalls, terms of service and robots.txt, the allegations in this case are that both OpenAI and Microsoft circumvented these restrictions in order to download data of news content.

JUDGE STEIN: Let me stop you there to make sure I understand this.

I issue opinions. Somehow those opinions end up either in the Westlaw or F. Supp., and I am told that each -- I don't know if this is true -- each night, certain services scrape the Southern District of New York for all opinions -- maybe memorandum, maybe orders, I don't know -- that have been issued today, and depending upon what their requirements are, let's say, scrape only opinions, not memorandum, not orders. Their computer will pick up that information from the Southern District of New York website and transport it to the owner of the service. Is that scraping?

MR. NELSON: Yes, that is, and Ms. Maisel will address it in more detail. But the short answer is yes. It's also a term of crawling, which is that these crawlers will go across the web. Then the scrapers will take the content. And by the way, one time is small. Sometimes these sites are hit hundreds of thousands a day by these crawlers and scrapers across the internet, and then it takes the data from those websites and then puts it on their own servers.

JUDGE STEIN: What's the hundreds of thousands times a day? What is that?

MS. MAISEL: Hello. Jennifer Maisel on behalf of the News Plaintiffs. Scraping is a big part of the news cases, so I will just jump in here.

JUDGE STEIN: OK.

MS. MAISEL: Just to contextualize this, the moment you connect something to the internet, lots of automated scripts begin pinging whatever it is you connected. There are great studies from years ago, if you plug in a toaster, a smart toaster, into a network, thousands of hits within the first hour, even the first moments that you plug something in. So there's a lot of automated activity on the internet.

JUDGE STEIN: I don't understand. What's the thousands of hits?

MS. MAISEL: So different independent scripts -- so from a technical perspective --

JUDGE STEIN: By the way, if none of this matters to the case, just tell me that.

MS. MAISEL: It does matter. So a bot will hit the server, which is where the content is hosted. That is how -- from a web browser, from a mobile device, whatever the device is that you are accessing content, your device will receive that content from a server where the content is hosted. As a user, you can control that behavior through your web browser,

browsing the internet, using an application, but there are automated scripts that will ping the server. They will send a request to the server, say, Send me this content, send me this website file, and then will download and scrape that content and store it in a separate database.

JUDGE STEIN: All right. So that's Westlaw pinging the Southern District of New York website, nysd.gov, and taking from it whatever its requirements are; all opinions, all -- everything, all orders, whatever it may be. Is that it?

MS. MAISEL: That's correct.

JUDGE STEIN: Then they store them.

MS. MAISEL: Then they store it.

And you asked about restrictions. If we are talking about government sites, I will give you a different example; the U.S. Patent Office site. They allow users, they allow bots, they allow companies to scrape content from their site through a specific dedicated interface. It's called an application programming interface.

JUDGE STEIN: That's the API that -- in addition to consumer data, that goes to the preservation order. There is a -- the API.

MS. MAISEL: Yes.

JUDGE STEIN: Go ahead.

MS. MAISEL: That allows automated access to content. So if someone wanted to bolt down the Court's decisions -- I am

not sure if the U.S. Court website has an API, but the Patent Office does. You can go to that API, download troves of patent applications, public information, but there are restrictions on that. These automated bots, they are not allowed to go to parts of the website that are not available to the public.

MR. NELSON: We are going to skip ahead a little bit. We will come back to this. Let's go to Slide 27, please.

JUDGE STEIN: Thank you. I remember you were helpful in the last oral argument. That was very helpful.

MR. NELSON: This is just to follow up on this particular point. For news articles -- this is an article from the Center for Investigative Reporting. And usually they are in something called HTML format. And what the article that -- this is how it would be displayed to you in a web browser, but underlying that display is a bunch of HTML code. The HTML files that OpenAI and Microsoft downloaded will contain the expression they want for model training that will also include the other code that you are seeing on the screen, such as how to display, copyright management information, title, author, copyright notice, terms of use, and that's what's excerpted here.

And OpenAI and Microsoft, as we will discuss, don't want their model spitting out a bunch of random HTML or copyright terms of service, as that would allow their people to know that their work is on training data. So to isolate the

expression and hide the source, OpenAI and Microsoft will use computer programs called content extractors, examples of which your Honor referenced like Dragnet and Newspaper in the motion to dismiss ruling in the News cases. The ideal content extractor will strip out the HTML and CMI, leaving only the expression behind. That expression is then saved into plain text files which can be used for training or other purposes. A similar process is performed on books --

JUDGE STEIN: For my purposes, this is the point at which CMI is not sent along on the way.

MR. NELSON: That's correct, your Honor.

JUDGE STEIN: OK.

MR. NELSON: That's correct.

JUDGE STEIN: Let's go back to where you were with Source Number 2, web scraping.

MR. NELSON: Yes. And so sometimes scrapers will scrape directly from a publisher's website, but LLM companies also obtain significant parts of their scraped data, including news and books data, from a third party called Common Crawl. Common Crawl itself scraped much of -- I am on Slide 19, your Honor.

THE COURT: Yes.

MR. NELSON: Common Crawl itself scraped much of the internet in ways which violated paywalls, robots.txt, and terms of service, and made those scrapes available in a central

repository where companies like OpenAI downloaded them. As OpenAI has grown, OpenAI has licensed some content, including entering into licenses for training data with a wide variety of news organizations, including The Atlantic, The Associated Press, and The Financial Times.

For its part, Microsoft has recently licensed books from Harper Collins for the very purpose of AI training, paying $5,000 a book for a three-year license to use the book in AI training. As we mentioned above, all of this activity, both illicit and licensed, is because getting a lot of high quality data fast is important to creating competitive LLMs and staying ahead of the pack. In real estate it's location, location, location.

JUDGE STEIN: What is high quality?

MR. NELSON: We are about to discuss that, but it is --

JUDGE STEIN: Don't let me --

MR. NELSON: No, no, your Honor.

JUDGE STEIN: I don't want to take up your time.

MR. NELSON: This is an incredibly important point. High quality means, basically, to quote something that is now public because of the anthropic decision, content that an editor would approve of, something that has gone through layers of editing that would allow -- it is not just low quality internet junk. It is something that has gone through a process

that reflects quality human expression.

So recall the OpenAI researcher I quoted at the start. The "it" in AI models is the data set. The data set is only as good as the inputs, and the ability of an LLM to convincingly mimic human expression is a function of the quality of the data used to train it. If you want your LLM, to answer your question, to emulate well-structured text of expressive variety, it needs to see well-structured text of expressive variety during training, and a lot of it. And if you want your LLM to produce text that is written in modern vernacular, and that incorporates modern developments, you will need in-copyright books and news. Books and news articles are an especially valuable source for the training set, as study after study has confirmed. This is the next slide.

For example, the creators of a pirated books data set called Books3 wrote in a 2020 paper introducing that data set that the pirated books were included because, quote, "Books are invaluable for long-range context modeling research and coherent storytelling." By the way, it was called Books3 because it was meant to mimic two of OpenAI's pirated data sets known as Books1 and Books2.

Similarly, researchers from Princeton noted in a 2024 paper that books are especially important because they provide long context data crucial for long context performance of the models. News data is similarly prized by AI companies. When

OpenAI trained one of its first LLMs called GPT-2, it created a data set called WebText which emphasized document quality. OpenAI released top domains in that data set, and as this example from that list shows, OpenAI's emphasis on document quality resulted in a list full of news content. OpenAI turned to the quality books and news data when faced with the challenge of filtering these huge web scrapes with low quality data.

From the paper announcing the release of a key OpenAI LLM called GPT-3 in 2020, OpenAI researchers wrote, that in order to filter the low quality internet documents which dominated these scrapes, they used the original WebText as a proxy for high quality documents and trained a classifier to prioritize documents which were predicted by the classifier to be higher quality. And OpenAI went on to note that it "used a collection of WebText, Wikipedia, and our web books corpus as the positive examples," and for the negative examples, they used unfiltered Common Crawl.

Now, by the way, the web books corpus was a euphemism. It actually comprised books from LibGen. This 2020 paper, incidentally, was seen as groundbreaking and helped establish OpenAI as a leader in the field.

JUDGE STEIN: Let's go back to 24. I have seen this chart before. What does it mean, domain count for The New York Times being the highest? New York Times pointed that out

previously.

MR. NELSON: Yes, your Honor. It's from -- and Ms. Maisel, I am sure, will correct me if I am wrong, but it's from Reddit, basically. So these are sources that were gathered from Reddit, and so it's the number of links that they used Reddit as a source for what was quality data because it was, again, curated by humans. And so these were the links that came from those Reddit downloads that they used. And so you can see how The New York Times dominates that list --

JUDGE STEIN: Is this where people were downloading information from Reddit, they did it 333,000 times from The Times and 44,000 from New York Daily News?

MR. NELSON: No, it's not people, it's OpenAI itself that downloaded Reddit copies.

Is that right, Ms. Maisel?

MS. MAISEL: We can think about these domain counts as articles, so individual web pages that were scraped and included as part of this data set.

THE COURT: And where is the data set?

MS. MAISEL: The data set is called WebText. It's described in the papers shown on the slide.

MR. NELSON: Let's go to the next slide. It's an OpenAI data set, your Honor.

JUDGE STEIN: OK.

MR. NELSON: And so once this information, these books

and news articles are downloaded, Defendants store huge quantities of material in centralized repositories. By the way, there are some -- in addition to OpenAI data sets that contain some of this information, there are also, your Honor, I would say, publicly-sourced open web data sets that anybody can access as well. We see that, for example, in what's called the pile, which is part of what Books3 is composed of. And certainly during this 2017 to 2019 time frame, there were these publicly-accessible data sets as well.

But the articles are downloaded in a variety of file formats like HTML, PDF, ePub and MOBI. The AI companies want to take that raw material and those disparate file formats, discard anything that isn't valuable protected expression, and hide the source of the training data by stripping out the copyright management information.

We discussed all this, so we will skip ahead to Slide 29. With the data acquired --

JUDGE STEIN: Let me just look at it.

Yes, sir.

MR. NELSON: Thank you, your Honor.

With the data acquired, copyright information stripped, and the raw material rendered into plain text files, the pre-training process can begin, and that's what we will address. At a high level, pre-training involves taking the data acquired at Step 1 and chopping it up into word fragments

known as tokens. It's represented by a number. The model then takes those tokens, which are, again, numerical representatives or words or word fragments, and matches them to a long series of numbers called vectors.

Vectors act like latitude and longitude coordinates to help locate the token relative to other tokens. The vectors encode information about the token and about the surrounding context in which the token appears. The model then predicts what the next token will be based upon that vector information, and millions or billions of iterations. The mathematical heart of this process is a measurement known as a loss function. The loss function measures how much the model's predicted output diverges from the actual training data. The model training process optimizes to reduce the loss measured by this loss function. By optimizing the model to reduce loss, the training process necessarily works to increase the probability that the model will reproduce the training data.

At the end of pre-training and after all of those iterations, you have a pre-trained model which is called a next token predictor, meaning it is literally predicting what the next token will be. The model's predictions about that next word are entirely dependent on the expression it encoded from its training data and the input provided by the user.

So how does an LLM encode the human expression in books and news articles? A helpful framework to understand the

basics of the LLM training process is a paper from Google AI researchers which said "language modeling is compression." As they put it, "It has long been established that predictive models can be transformed into lossless compressors and vice versa." Think of a zip file. If you have a document that you want to send via e-mail, you might compress it into a zip file to reduce its size. The way this works is that the document, in this example a doc file, has a bunch of material that is common to all doc files. The zip program has hard-coded rules that will remove that repetitive material and thus reduce the size of the file.

In other words, compression takes advantage of these redundancies in data and converts them to shorthand. When the recipient receives the compressed file, they can then use the zip software on their computer to uncompress the file, and the software will add back in those portions that were previously removed. At the end, you have exactly the same data that was compressed at the start.

As the Google researchers put it, "current language model training protocols use a maximum compression objective." In short, both LLMs and more traditional compression algorithms are optimized using loss functions to reduce the size of the data in question. Training a language model to predict what that next token is, is in a fundamental way the process of compressing the training data not a process of, quote,

"learning patterns of language," as Defendants say and you will see in their presentation. It is technically incorrect in a literal sense to say that the model is learning. Instead, the model training process is a massive process of predicting the next token.

JUDGE STEIN: What's the difference?

MR. NELSON: Because it is literally just predicting the next word.

JUDGE STEIN: I understand. Why can't that be characterized as the model is learning?

MR. NELSON: Because the model is not human.

JUDGE STEIN: Right.

MR. NELSON: But the model is taking in its input and it is doing mathematical algorithms to just figure out what that next word is. So, for example -- so maybe it's just more of a semantic point, your Honor, but it is a mathematical process to predict the next word.

JUDGE STEIN: That I understand. And I understand how the idea is to decrease the loss ratio, in other words, to make the computer smarter, that is, to have it predict what actually would be that next token. So the smaller the loss amount, the more efficient the machine is. Am I right so far?

MR. NELSON: Yes. And let's just go quickly to the next slide. I will skip this as a matter of time, but you can see that's exactly right about how the models are trying to get

as close as possible to a -- what's called test loss on this slide -- this is an OpenAI paper -- as close to zero as they can to approximate the data set.

JUDGE STEIN: And again, I apologize for slowing you down. I will give you a little more time because I am holding you up.

I understand the predictive nature of what you are doing, and I understand how a zip file somehow compresses the information. I think there was somebody who figured out how to be able to send sports -- basketball games, and so forth, by computer, and it involved a lot of compression. So I understand the predictive issue.

Why is compression important, and how is it relevant to this case?

MR. NELSON: Well, it's important because what it means is it is -- in a fundamental way, the model is its inputs.

JUDGE STEIN: Right.

MR. NELSON: That's what it is. It is not -- so we are going to hear from the other side, as Ms. Maisel will also get into, that it can do a lot of things based upon the inputs, but that is irrelevant to what the model contains. The model is, by design, exactly supposed to represent what its inputs were.

JUDGE STEIN: Right.

MR. NELSON: And so that is why compression is -- in other words, there are -- it's not like -- there are certainly mathematical computations going on in the model, but at its core, what is trying to go on is that the model is the training data, and that's what the training process optimize is for.

So, for example, let's go to Slide 36. Take "One small step for man, one giant leap for," blank. The model might predict penguins, mankind, or frogs. We know that if it picks penguins it's wrong. It checks to see that mankind is the right answer, updating it the next time. This process happens millions or billions of times to train a single model. And each time at the prediction phase, the model is checking its prediction against that underlying text in its training data for accuracy and to improve the compression, to reduce the loss.

JUDGE STEIN: The model is asking, "What's the next word?" and it will make a prediction, and it then checks its prediction against the famous quote, and it will find out that mankind is the, quote, "correct" prediction?

MR. NELSON: That's exactly right, your Honor. And that will happen -- so it will get better and better, it will happen literally millions, billions of times.

JUDGE STEIN: So it needs your high quality data to underlie its data set.

MR. NELSON: That's exactly right. And, in fact, if

the model process is functioning properly, the model should continuously get closer and closer to a perfect compression.

JUDGE STEIN: But it can't do that without having the data set; is that right?

MR. NELSON: That's right. It needs the data set to have the perfect -- that's why going back -- what are the two factors of production? You need the chips to make it work, but you need the input of the data. And without the input of the data, you have nothing, OK.

JUDGE STEIN: If the underlying data has some low quality document that says, "One small step for man, one giant leap for frogs," it will register that.

MR. NELSON: That's correct. That's correct. And so how it learns, in their parlance, OK, is how it takes in information, OK. And so the -- that is why the high quality data is critical to making it function, because if you are on the output side, and you say, "How does the quote end, 'One small step for man, one giant leap for frogs,'" you would think that something is wrong with the model, right.

And again, you don't have to take our word for it or even Google's about what is going on with respect to compression and prediction. Here again is OpenAI cofounder and chief scientist, Ilya Sutskever, on the fundamental equivalence of compression and prediction in the context of LLMs: Quote, "It is well-known that compression is prediction. Every

compressor can be a predictor and vice versa. There is a one-to-one correspondence between all compressors and all predictors."

To translate, he is saying that during the LLM training process, the model is adjusted continuously by measuring how closely the model reproduces its training data. The goal of the process is to optimize the model towards a more perfect approximation of its training data, and that process is repeated over and over, millions, billions of times, until the model approximates a compression of the training data as measured by the loss function.

JUDGE STEIN: That I understand. But when it says "every compressor can be a predictor," what is the compressor when it says "every compressor"?

MR. NELSON: They are talking about the model. They are talking about --

JUDGE STEIN: The data set?

MR. NELSON: Well, no, the end model, right. So the model is composed of the data sets, so they are talking about the model itself. And so the compression is seen not only in the writings in public statements of the leading AI researchers, but in the vast -- we know this because of the vast amount of memorization which the models exhibit. In fact, in OpenAI's seminal 2020 paper on GPT-3, they said, quote, "that large models have the capacity to memorize vast amounts

of content."

Ms. Maisel very shortly will talk more about memorization, but I want to be clear that we are not claiming that every model memorizes every piece of training data, but that on a technical level, the rampant memorization which these models do exhibit is a necessary consequence of the nature of the training process, and OpenAI and Microsoft's goal in optimizing their LLM is to minimize the loss. At its core -- then I am going to turn it over to Ms. Maisel -- the LLM training process aims to produce a model that emulates the human expression present in its training data. And Ms. Maisel will pick it up from there, after the pre-training process.

JUDGE STEIN: So the bottom line of your presentation is the models are designed ultimately to predict what is in the training data.

MR. NELSON: That is correct, your Honor.

JUDGE STEIN: OK. Got it.

MR. NELSON: Thank you, your Honor.

THE COURT: That's why you need "high quality stuff."

MR. NELSON: That's why you need high quality stuff. That's exactly right, your Honor.

JUDGE STEIN: OK. Thank you.

MS. MAISEL: Jennifer Maisel again on behalf of the News Plaintiffs.

So to pick up where Justin Nelson left off, I am going

to talk a little bit about mid-training and post-training. And your Honor hit on an important point earlier about the frogs example; what if you had low quality data that had frogs at the end of it? Is the model going to pick up on that? And one reason the model may not pick up on that can be answered with this mid-training and post-training process. And there are two points I want to hit on mid-training and post-training. After pre-training, it's important to consider the order in which data is fed into the training algorithm; this loss function that trains the model to accurately reproduce its training data to predict the next token in its training data.

High quality content is seen later in the training process, and it's seen later in the training process for the important reason that the model is more likely to encode and retain the information from that high quality data that it has last seen. So the ordering is important.

The second point that's important is there is a lot of low quality data out there, especially if we are talking about scraping the internet. Tons of low quality stuff. And you will see a slide, at least from what we saw last night from OpenAI, just showing the shear volume of content that it needs to train these models. There isn't enough data in the world to only use high quality data to train the model. So some of that lower quality content seen earlier in training to get general -- get a general idea about how words are put together

and learn about the English language, learn about other languages, get some structure underway. But high quality content, the books, the news content, high quality data sources, this is seen later, and it's reinforced as part of this timing aspect. So if we talk about mid-training as well as post-training, having reinforcement learning, these are different algorithms, but different algorithms to help steer the model towards this higher quality data, these higher quality responses.

And a few months ago I spoke with you about some of the technical issues in the case. I called it ingestion, digestion, and regurgitation. And I think these concepts are helpful to just understand a little bit more why these models tend to memorize their training data; memorization being the act of encoding some of the training data within the parameters of the model itself. And we can see evidence of that in the form of regurgitation when we see the model actually output verbatim copies of its training data.

So if we go to Slide 42, the problem with the pre-training data sets, as I mentioned, is there is a massive quantity of low quality data, misinformation, bias, toxic information, but we have to deal with that kind of data because we don't have enough high quality data.

The mid-training process involves curating a select portion, probably a subportion of that pre-training data, and

identifying high quality data to use later in this training process as part of mid-training or as part of fine-tuning. The post-training process is designed to teach the model to excel at specific tasks. And there is a lot of human involvement here. The developers, the OpenAI, Microsoft, the people that are building these models and training these models, they want to train them to do specific tasks really well. In the case of an LLM coupled with a search system, we want that LLM design to be able to summarize search results really well. In the case of a chat bot, we want the chat bot to be trained to engage in conversation in a friendly and helpful tune. So post-training helps steer the model in different directions, and as part of that post-training process, we are steering the model again towards that high quality data, books and news.

So as I mentioned, high quality mid-training data sets, it's a subportion of the pre-training data. And there are a lot of different uses that I won't get into here, but in terms of identifying these high quality data sets, high quality data sets may be downloaded and used just to identify those high quality data sets. Mr. Nelson referenced the GPT paper, the GPT-3 paper using WebText and other high quality data to identify other high quality data.

(Continued on next page)

MS. MAISEL: So, there are lots of different uses of content and high-quality copyrighted content as part of the training process. Again, I won't get into it here, but the point here is that we are selecting and curating high-quality data that includes news and books content.

The other objective of mid-training is to help update the model with new information. Back in January, your Honor asked me if these models are continuously learning, and one answer is yes. There are two days to do that, it's part of this mid-training and post-training process. Mid-training, if you, for example, collect a set of more recent news articles and more recent news content, you can update the model, you can train it again so that it's able to emulate the new expression it has seen from the news content.

Similarly, you can update the model with specialized domain expertise. Several of the News plaintiffs publish high-quality expert and informative content. I have up on the screen a lot of the Ziff Davis brands. Very unique, specialized content. And if you want the model to be really great at a specific task, you're going to use that specific content later in the training process again so that --

THE COURT: These are Ziff Davis products, is that it?

MS. MAISEL: That's right. They're Ziff Davis prints, Mashable, BabyCenter, PCMagazine. Several of these brands are in the lawsuit and have very unique high-quality content in

different domains and specialties.

And then similarly, high-quality books content is also seen in the mid-training process. For all the reasons Mr. Nelson explained, books have a lot of value to LLMs. They're very long, they're well written, they have a diverse set of language. It's great for these models to see them later in the training process to learn from them and to enhance the capabilities of the model.

So, post-training. Post-training encompasses a lot of different techniques. We don't have time to get into the weeds on all of them, but as I said, we're training the model to excel at specific tasks. And this is an iterative process. There's a loop shown on the screen. This is from OpenAI's website.

THE COURT: Before you go on, go back briefly to slide 42 where you started this pre-training, mid-training, post-training. It seems to me that that's all one process, that there's really no difference between one, two, and three except presumably the model is getting more predictive and more predictive as more data is fed into it. Isn't that all that one, two, and three represent?

MS. MAISEL: Yes. These are all part of that ingestion process, that training process. There are different nuances to each of these phases, including the differences in the datasets used.

THE COURT: You've got a huge amount to begin with and then you narrow it down by getting specific high-quality data in, and then that continues on with not only specific high-quality data but stuff designed for specific purposes, but it's all a continuum, it seems to me.

MS. MAISEL: It is all a continuum. And even in the pre-training process -- it's a continuum in the sense that things can go back and forth, we could go mid-training, post-training, back and forth. And even in pre-training, a lot of the higher-quality content tends to show up later in the pre-training process. So, these are all interrelated concepts.

THE COURT: Okay. Go ahead.

MS. MAISEL: So, in post-training, as I said, it encompasses several different techniques, and these are techniques designed and identified by the developers of these systems. What do they want their LLMs to do and be capable of doing. And these datasets tend to include example conversations, examples of model prompts and examples of model outputs.

And one example, if we go to slide 48. If we have a search-based product, we might want to train the LLM on a set of conversation data that exemplifies how the model can search the web. You know, what kind of queries the model should send, what kind of responses the models should send, and have the model just learn low to browse the web, receive web results,

and present those web results to users, so we might have a specific conversation dataset devoted to web browsing.

And then finally, we have reinforcement learning through human feedback. I just point this out because it has been in the news. There's one article shown on this slide about some of the human reviewers, the preverbal man behind the curtain. There's lots of human input as part of the training process, and one goal of these human reviewers is to help steer the model responses away from toxic data, away from offensive responses, misinformation, answers that are just not going to be helpful and useful in commercial products like we have here, and instead steer the model responses towards high-quality content.

So, just to recap, after pre-training we have what's called a base model, and these mid-training and post-training sets, they help to form the final model that's ready for productization.

And I talked about memorization. If we're tying it back to our uses here and what copyright infringement uses we're referring to, they're different theories: Memorization is evidence of the large language model as containing or encoding a copy of its training data, so copies of copyrighted news and books content. And another piece of evidence of that memorization is output, regurgitations.

And as we've just walked through, high-quality content

is more likely to be memorized, and those mid-training and post-training phases help explain why that phenomena is true. There's a higher likelihood of memorization of content seen later in training, and we also have alignment of the model to mimic the high-quality content that people want to see in the outputs from the models.

THE COURT: What is memorization? Isn't memorization simply a zero loss factor? In other words, the token that the model predicts will be 100 percent accurate of the underlying data.

MS. MAISEL: That's correct.

THE COURT: Okay. Got it.

MS. MAISEL: And it may not be 100 percent, it may be a little bit off, and we're going to talk about hallucinations. But generally speaking, yes, it's to generate an accurate or substantially similar output to the original --

THE COURT: Okay. It's not really memorization, it's 100 percent accurate predictive quality.

MS. MAISEL: Yes.

THE COURT: Okay.

MS. MAISEL: And as you have seen, as been discussed extensively in Exhibit J to The New York Times complaint and the Daily News complaint, we were able to obtain --

THE COURT: I thought I was done with that. I was hoping I was done with it.

Go ahead.

MS. MAISEL: Again, this is an example of memorization of the models. We were able to get 100 examples of verbatim regurgitation from the models, evidencing this memorization process and the high emphasis that the defendants use to train on used content.

All right. Let's talk about the output case. I'd like to skip ahead to retrieval augmented generation, just because I think we could use some clarity there.

Another problem with training is that it takes a lot of time to complete. As Mr. Nelson said, there's a lot of computing resources required, a lot of data that has to be processed, trillions of iterations for the training algorithms, and therefore there are gaps in the expression that LLM can produce from its training data. And the solution is to augment the LLM with real-time, up-to-date content taken from the web.

We have up on the screen, for example, an LLM that's out there in the market. It's not going to be trained on the fact that yesterday there was a heat wave across many parts of the country, so how do we update the models to ground them in this up-to-date information? The answer is retrieval augmented generation.

I'm going to walk through the basic steps. At step one, a user -- so an end user like myself -- I might submit a prompt to an LLM maybe through a chatbot, maybe through an API.

There are different interfaces. But I'll submit a prompt. And the LLM will then translate that prompt to a query to an external data source. That external data source could be a website on the internet, it could be a database, it could be any external source, and that data source is going to return a copy of content contained in that data source.

For example, if we're accessing a website, the copy returned as part of that response might be the text taken from that website. Mr. Nelson talked about the scraping aspects. We might scrape article content from a website and return that back to the LLM, and the LLM is going to use that scraped content, along with the original prompt, to formulate a response.

And entities out there have noted that the retrieval augmented generation traffic is growing exponentially and is even exceeding the rate of growth as compared to all of the bots out there that are used for scraping, indicating that RAG is a very important use case for the LLM's very important future.

ChatGPT includes many features that rely on RAG. I won't get into all of these, but the Browse with Bing, SearchGPTs like Remove Paywall, all of these features within ChatGPT rely on retrieval augmented generation to go out, fetch external data sources, and return them.

Now I want to spend a moment just talking about

Copilot, which is Microsoft's retrieval augmented generation chatbot.

THE COURT: Let me try to synthesize what retrieval augmented generation is and tell me if it's correct. All it is, is the LLMs getting better in their predictive quality by incorporating the prompt of the end user, you or me, and the response of the LLM back to that prompt. Is that all RAG is?

MS. MAISEL: Yes and no. Yes, in the sense that we're augmenting the prompt. So, let me give you an example. If I want to know today's headlines, I can ask a chatbot, you know, what has The New York Times reported today. Again, hypothetically. And the chatbot might go out, using retrieval augmented generation, to The New York Times' website, pull out the latest headlines, and formulate a new prompt to provide to the LLM. That prompt might say, user asked this, they asked what today's headlines are. Formulate your response using all of these headlines that I have just taken from The New York Times' website, and then the LLM will generate a response using that content.

THE COURT: Okay. That's a little different. Thank you.

MS. MAISEL: Yeah. And that additional content, that's what we call grounding. So, grounding the LLM response with that additional content.

I just want to take a moment to talk about Microsoft's

Copilot chatbot. We saw Microsoft's presentation last night, and there is a dispute about whether the current iteration of Copilot out there in the market is the same or was alleged of infringement in the original complaints. And two things I want to highlight for your Honors:

One, the current iteration of Copilot, which was released on October 1, 2024, it uses the GPT models, the same GPT models that the prior version of Copilot used, and it uses the same retrieval augmented generation process insofar as it's relevant for the claims at issue.

The wrinkle with Copilot is instead of accessing any website or any specific data source, it's going rely on the Bing Search Index. This is Microsoft's Bing search engines, the search index that that Microsoft Bing search engine uses in order to formulate search results.

So, that's the difference: Same functionality as the original Copilot alleged of infringement in the complaint, same GPT models, same infringing retrieval augmented generation technology. That's it. Any of these other changes that you'll hear soon enough are totally behind the scenes.

So, skipping ahead, I'm just going to talk about the types of products. There are three buckets, I would say: One is the application programming interface, second are chatbots, and third are chatbots with RAG capabilities.

And the API, we talked a little bit about that

earlier. And API is a mechanism that allows people and programs, automated computer scripts, to directly access the GPT models. They're interfaces that OpenAI provides. We can show that on slide 65 or skip ahead.

This is an interface for the API platform. It's a friendly human interface, so you can log on, you can go to the API platform that OpenAI has available, and you can submit one-off prompts similar to what you would do with ChatGPT.

The wrinkle here, though, is you can create an automated script or program to access the API, to access the models, and generate content at unprecedented speed and scale. So, instead of --

THE COURT: The API allows the end user to directly interface with what?

MS. MAISEL: With the GPT models.

THE COURT: Okay.

MS. MAISEL: So instead of intermediated through a chatbot or through some other service, direct access to the models.

And the problem with this, the ability to generate content at immense speed and scale, leads to the creation of a ton of AI-generated low-quality content out there. This low-quality content is affecting the news industry. As OpenAI recognized, and this is on slide 66, they knew that the API, when they first launched it back with GPT-3, that the API can

be used to create false, misleading, or propagandist essays, tweets, and news stories *de novo*. And with that initial launch of the API, they learned a lot of lessons. They said that they detected and stopped hundreds of actors attempting to misuse GPT, and that these actors were misusing the API in ways that they didn't anticipate or that they anticipated but did not expect to be so prevalent.

And we have a Rule 71(a) objection, your Honor, that's on the agenda for today. This API data, the ability to generate content at immense speed and scale, has all been destroyed. The records of the creation of content of this misinformation, the misuse of their API, virtually all of that data has been destroyed.

Skipping ahead, NewsGuard is another third party watchdog out there that is tracking AI-enabled misinformation. They've identified 1,271 different websites.

THE COURT: What does the disinformation stem from in the API as opposed to just consumer data, GPT?

MS. MAISEL: So, it's speed and scale. Someone can write an automated script saying, hey, write a bunch of biased news articles based on The New York Times or based on the Daily News or based on other publications, use that content as a starting point but bias it so it says something different, so that the content looks plausible or remotely plausible and can be used to target specific objectives and goals.

And one example al of this, the proliferation of AI generated news and just the proliferation of spam content out there. As we've seen in ChatGPT, if we go to slide 68, there are known plagiarizers out there, they've been out there for a while before the prevalence of these large language models.

THE COURT: Did you say "known plagiarizers?"

MS. MAISEL: Yes, known plagiarizers.

THE COURT: All right.

MS. MAISEL: One of them is this entity called D-N-Y-U-Z, DNyuz. And DNyuz is a known plagiarizer of The New York Times, and we see on the screen here ChatGPT citing DNyuz, a known plagiarizer, for content from a New York Times article instead of citing to the original New York Times source.

Another example of this is provided in the complaint, The New York Times complaint at paragraph 121. And this is a problem: The proliferation of AI-generated news and misinformation and low-quality AI-generated slop is crowding out publishers from the web, and these publishers are being driven out. Instead of citing to the original source, these chatbots and other products are instead citing to these plagiarizers, these other websites. So, there's this crowding out in the market.

And similar issues are impacting the books industry. I have up on slide 69 this point that the GPT models have also supercharged --

THE COURT: So, help me with exactly what the claim is if you're arguing that they're not duplicating New York Times articles. They're producing a lower-quality misinformation product. What's your claim?

MS. MAISEL: So, this inquiry --

THE COURT: Dilution of the value of The New York Times?

MS. MAISEL: That's exactly right. And if we're thinking about the fair use analysis, factor four, the substitution of the copyrighted content in the market, that original copyrighted content is getting crowded out by this AI-generated content.

THE COURT: All right. That's helpful. Always going back to the fair use factors helps me because as I've said, I think that's what's really important here, the fair use defense.

Go ahead.

MS. MAISEL: Yeah. So, in addition to news, we have this issue with books. In some instances, right before or during the release of a new book, online marketplaces are flooded with knockoff books that are designed to deceive customers into purchasing the sham book, thinking it is the original. AI sham books are frequently uploaded under the names of established authors, hoping to trade on their notoriety to steal sales from genuine works. Even where the

AI-generated works do not use the author's name, they're being uploaded in huge quantities to flood the supply of books and make it impossible to --

THE COURT: You have to speak a little more slowly. The reporter needs to take it down.

MS. MAISEL: No problem.

These sham books, which can be generated with the press of a button, then can receive royalty payments by programs like Amazon's Kindle Unlimited program.

These sham books have also appeared at huge volumes in the immediate aftermath of prominent deaths. I'm on slide 70 here. They contain misinformation and seek to profit off the surge and interest occasioned by celebrity death while crowding out legitimate authors. In one example, an AI-generated biography falsely described a recently deceased New York Times editor as a chain smoker who honed his skills in Cairo and reported from Vietnam, none of which is true. And just one of these AI-generated authors published multiple sham book a month, a rate impossible for a human author and which impedes the recognition of legitimate works.

So, I just want to spend a moment talking about Next Generation search. On the screen here you have some examples of news cases for ChatGPT taken directly from OpenAI's marketing material.

But if we skip ahead, if we go to slide 72, at our

past technology tutorial, your Honor had asked what is the difference between regular, traditional online search and the Next Generation search, as Microsoft's counsel put it, with the Copilot chatbot. And what Microsoft's counsel said is what is great about them is that they're actually designed to answer your question. So it's not just giving you a series of links you can go look for yourself and maybe find the answer.

So, the next slide we have just a side-by-side comparison of a response from the Copilot chatbot used for a search, and then traditional Bing search interface on the right. As you can see on the left, at the very bottom of the screen, very small font, there's a "learn more" statement with a couple of hyperlink links after it. It's very obscure hyperlinks at the bottom of the page.

By contract on the right, we have the traditional list of blue links. We have an excerpt here with a very prominent link to the article at the time, Lakers Fix Errors on Kobe Bryant's Statue, and a very small snippet from the article. And it begs the question, which link is a user more likely to click on, from the new Copilot interface or the original Bing search interface?

And to be clear, the defendants know this answer. I'm sure that they are tracking this information as part of monetizing their products and tracking engagement with their products, but we have some third-party reporting out there.

One of those entities, Tollbit, who is helping publishers, mediate this new AI space, they've noted that Tollbit has seen AI bots on average sending 95.7 percent less traffic than traditional Google search. So almost 96 percent of people are less likely to click on the links shown in these generative AI search products than in traditional search.

THE COURT: I'm sorry. And what does that mean on 73? They're more likely to lick on the links in the Copilot output than the Bing output?

MS. MAISEL: No. So, users are 96 percent less likely to click on the link in the Copilot output as compared to the Bing traditional search output.

THE COURT: All right.

MS. MAISEL: And again, on slide 75, we have more third-party data, people tracking what they can of traffic from these sites. And one of them is CEO of Cloudflare said, "People aren't following the footnotes." Namely, they aren't clicking on the links, and there's been a tremendous dropoff in traffic.

So final point on output. We talked a little bit about hallucinations at our first technology tutorial. This is a big --

THE COURT: I'm sorry to slow you down. I guess I'm having a little trouble with this last point of yours on 73. Perhaps because I can't possibly read the text on the left from

Copilot. Tell me again what you're saying.

MS. MAISEL: So, what I'm saying, on the screen on the left, which is --

THE COURT: The one on the left has a series of links at the very bottom. The one on the right, the traditional use, is rather straightforward. It says what the topic is and has a link to a single source. I see that.

Go ahead.

MS. MAISEL: Yes. So the point is, in traditional search, the hyperlink was shown at the top very prominently. You can see that on the right. It's in purple, bluish text there.

THE COURT: Yes.

MS. MAISEL: Very easy to click on. And the output from Bing search, it's just that two lines of text. Most of it's metadata, the publication date, and then one line: Los Angeles. The lakers on Tuesday finish fixing multiple errors on the Kobe Briant ...

Compared on the left for the same article from the Orange County Register, you have the entire article text reproduced in the Copilot response. If you just read through, the user asked: Please provide the full text from the Orange County Register articles.

And the Copilot response is: Certainly. Here's the full text from the Orange County Register article.

And then it goes on to provide the text from the article. It even includes a couple images. And then at the very bottom, after already providing the text to the article and what the article says, it says: You can click on these links and learn more.

So, the question is, will users click on these links if they're already being provided the content within the Copilot interface within these other chatbot interfaces.

THE COURT: So, your point is less traffic will go to The New York Times or, in this case, the Orange County Register.

MS. MAISEL: Exactly, your Honor.

THE COURT: Okay. Because more information is being force-fed to the end user. Or -- I won't load it that way. More information is being provided to the end user by Copilot than by Bing.

MS. MAISEL: Exactly. And to frame this within the fair use analysis, the word I would use is it substitutes. It supplants the original. People are getting the content within the chatbot interface --

THE COURT: And not going to Orange County Register and not paying Orange County Register.

MS. MAISEL: Exactly. When they'll hit that paywall or when they'll drive advertising or other revenue by going directly to the website.

THE COURT: All right.

MS. MAISEL: So, we talked a little bit about hallucinations back in January. We talked a little bit about memorization. Sometimes the model doesn't quite get it right or it misinterprets something. It doesn't misinterpret something but it's reproducing expression that's just not quite right. And we have examples of these so-called hallucinations throughout the complaints. And even in the presentation that we received last night from Microsoft, we found a little hallucination in there.

If we go to slide 79 of our slide deck, on page 15 of Microsoft's technical tutorial which has a video, a user was asking for a business plan for an eco-friendly café in Manhattan. And the Copilot response has that business plan, and section two of that business plan indicates local competition. And local competition included a coffee shop called Grounded Coffee House. And this was in 2025, this was a very recent answer. And as we can see on the right, Grounded Coffee House closed almost five years ago. It's not in business. So why is Copilot outputting Grounded Coffee House as a local competitor to consider as part of your business plan? It's a minor example but it's an example of a hallucination. It's an example of output that's not quite right. And had this been citing a news source for that proposition, it tarnishes that news source.

MAGISTRATE JUDGE WANG: Is this suggesting that retrieval augmented generation, or RAG, can't account for lack of information if it goes to retrieve?

MS. MAISEL: So, one thing that you will hear is that RAG and grounding will help make the responses more accurate because it's pulling content directly from the source and using that content copied from that source to provide a model output. And sometimes it still gets it wrong.

Here, wherever Copilot pulled this information, this coffee shop had closed and the model just didn't pick up on it. It picked up some old source that said, hey, this might be a competitor.

And just one final point here while we're on this slide. Again, in the presentation Microsoft shared last night, this behavior is exhibited by the current version of Copilot, or at least what appears to be the current version of Copilot, that Microsoft contends is not in the case. So, as we'll go through Microsoft's presentation, a lot of what we're going to see is from the new product as they allege that apparently is not in the case but does the exact same thing as the old product.

And then finally, there's one final point, again, while we're talking about outputs in this hallucination issue. Another point that you're going to hear is that the LLMs are creative, they come up with different responses to the same

prompt. And I want to highlight that. That creativity is more like pseudorandomness. Every time you put data through an LLM and get that next token prediction, the number of options -- so before we had mankind, frog, and another option, and different probabilities for that output, those probabilities will remain the same every single time you run that string of texts through the model to predict the next token.

But you can inject randomness into that process, pseudorandomness, computer randomness. If you have, say, the top three responses, instead of selecting the most probable response each time, you could instead instruct the model to select one of several different responses having a sufficient confidence value, a sufficient probability. So you can inject randomness by instead of selecting the most likely next token, instead selecting one of maybe the top five next tokens. And this feature, again, we're back on the API platform if we go to slide 65. If we're thing about temperature of the model, temperature and the top P values here, they're proxies for the pseudorandomness. If you turn the temperature down, you're making the output less random, you're selecting the most probable or more probable response, and you're going to get more predictable output each time.

So, for example, for the Exhibit J examples in the complaint, we wanted more accurate LLM responses so we turned the temperature down. Let's see what it trained on and if it

can accurately reproduce what it trained on. If you wanted a more creative response, so to speak, if we wanted to inject more randomness in how the model was selecting those output tokens, we could turn the temperature up and have the model select different tokens which will then lead the output in different selections based on that next selection of the token.

THE COURT: What does it mean to turn the temperature up or down? Are you talking literally?

MS. MAISEL: Literally. So, on the screen here is the API interface. And to be clear this is a feature available in the API. To my knowledge, you can't change these parameters in ChatGPT or Copilot or the other top bots out there. But in the API, you can customize the type of response you're getting.

So, shown on the screen there's actually a temperature option that you can open and there's a sliding bar there and you can make the temperature lower or higher depending on what kind of response you want. You also have other customization options. You could decide what kind of format you want, if you want text, images, other types of media.

There's max tokens, which is how long of a response do you want. Do you want to cap off the response at a certain length? And top P, roughly, it's from what selection of tokens are we selecting from. You know, are we selecting from tokens having a minimum or maximum probability value based on that prediction?

THE COURT: Put aside the API for just a moment. Go back to the beginning. If the object of training the LLMs is to get the loss factor as low as possible so that it accurately predicts what the next token is -- again, put API to the side, you can fold it in in a moment -- what accounts for the hallucinations, the incorrect answers?

MS. MAISEL: It's just sometimes it gets it wrong. So, for example, if you put in a prompt that the LLM --

THE COURT: What you're saying is even though the loss percentage goes down and down, there's still some factor in there and that can account for incorrect answers. Is that essentially it?

MS. MAISEL: In a sense. I don't want to get too far into it because a lot of this we just don't know. We don't know why some percent of the time the model will just --

THE COURT: I think Microsoft and OpenAI have said they're trying to deal with the issue of hallucinations and I think they've said they're not quite from where it's from either. Does the API change it in some way? Change what we're talking about in terms of the likelihood of hallucinations? I take it to the extent you're changing the temperature, it would.

MS. MAISEL: Potentially. So, if we want a more accurate, faithful response to the original expression that we are trying to derive from, turning the temperature down will

help ensure that that output more closely mirrors something from the training data, something that the model has seen.

THE COURT: All right. Thank you.

MS. MAISEL: Yeah. The hallucination issue, I'm not sure if any of us can explain why it happens precisely.

THE COURT: Is hallucination the same thing as incorrect answers?

MS. MAISEL: Yes.

THE COURT: Okay.

MS. MAISEL: And sometimes those incorrect answers look plausible on their face because they're --

THE COURT: But for example, the hallucinations that have been in the news in terms of the law are cases that simply don't exist. That's not a misplaced word, that's, like, crazy.

MS. MAISEL: Yes.

THE COURT: But for you, the misplaced word as well as the crazy thing is a hallucination; is that right?

MS. MAISEL: Yes, that's right.

THE COURT: Okay.

MS. MAISEL: And it's problematic when that hallucination then attributes one of our clients, when it attributes The New York Times, as the owner of that hallucination.

THE COURT: It hurts you.

MS. MAISEL: And that is the basis of our trademark

dilution claims.

THE COURT: Okay. Why don't you finish up.

MS. MAISEL: Yep.

So, one final point. I'm going to just talk one final point on outputs. This is slide 77. One point I want to stress is that there are certain guardrails, and I put them in quotes here because the defendants have implemented specific mitigation measures over time to try and prevent this phenomena of regurgitation and the ability to collect evidence of what they trained on. And these mitigation measures, they've changed over time. There's a slide I did not get through, but we have a timeline of the development of the models.

The defendants have been iterating on these models over time. They have made updates, they've made updates to the products. And what we've learned in discovery is that over time, the defendants have implemented additional measures to prevent or block the regurgitation of the content of the plaintiffs that have sued them. Some of these guardrails have come up post-litigation, some of them are pre-litigation but not always implemented at the time of the initial release of the product. And these mitigation measures were specifically targeted to copyright owners who I believe the defendants most likely thought would sue them. So, entities that may have been complaining about their works used for training without any compensation.

So, if we look at the behavior of the products today, even compared to a year ago or year and a half ago, they appear to look different because of these post-litigation, these litigation-inspired guardrails that have been put up.

THE COURT: All right. Now we're talking preservation issues. But if I understand you correctly, isn't it better for you if the LLMs are not spitting out The New York Times' articles?

MS. MAISEL: It's a small step, especially on things like hallucination. But earlier today I talked about the proliferation of these AI-generated news articles, the proliferation of AI-generated books. The fact that the models may not spit out the exact verbatim expression of the article doesn't mean that they're not spitting out summaries or these lower-quality reproductions of works that have the effect of substituting for these works in the market. So, it is still a problem, even if they eliminate or try to eliminate the problem of verbatim or substantial similarity in the output.

THE COURT: All right.

MS. MAISEL: And as we have seen, it still happens, so they're not perfect.

THE COURT: Thank you.

MS. MAISEL: Thank you, your Honor.

THE COURT: All right. Let's do this. It's 20 of. We've been going for an hour and a half. Let's take a brief

10-minute break and then I'll hear from the defendants. We'll go to about 1:00, we'll break for lunch, and then I want, for those who want to stay, an argument on the Rule 72 appeal from the preservation order. If we can fit that in before lunch, so much to the better.

Thank you. Let's take ten minutes.

(Recess)

(In open court)

THE COURT: Let me hear from OpenAI and Microsoft.

MR. VAN NEST: Your Honor, Mr. Bayley is going to present the OpenAI's tutorial. Would your Honors like written copies? We have them if you wish.

THE COURT: Yes. And I appreciate the News plaintiffs' presentation. It was very helpful.

Mr. Bayley.

MR. BAYLEY: Yes. Good morning, your Honor.

So, my name is Ed Bayley and I represent OpenAI, which is an AI research lab that was founded in 2015.

Now, what your Honor just heard, I would have to say, is a pretty highly argumentive presentation about a lot of disputed facts and evidence and characterizations in this case, but what was noticeably missing, I thought, was an actual cogent explanation of how this technology actually works and what you can do with this technology, which I understood to be the purpose of this tutorial, so that's what I intend to do for

the remainder of this morning and early afternoon.

So, I had the pleasure of telling you about OpenAI's groundbreaking, innovative technology which it calls the generative pretrained transformer or GPT. So what is a GPT? In 2018, OpenAI introduced the generative pretrained transformer as an innovated AI advancement capable of generating human-like text. By 2022, just four years later, this architecture was used to power ChatGPT, an AI chatbot that rapidly gained global adoption. ChatGPT has revolutionized workflows for both ordinary users and businesses alike, and today, hundreds of millions of people around the world use this transformative technology in countless ways. From the very ordinary like drafting an email to the sublime like helping the visually impaired navigate their life more independently.

So, I want to start, before I get under the hood, with a couple of examples of what this technology can do. So let's start with the ordinary. Here's the ChatGPT interface. It's simple but it's actually very powerful. You begin by entering a prompt to ask it to do something, so that's what I'm going to do. I'm going to provide it a brief prompt and it's going to ask: Provide a brief engaging description of the grading system at Yale Law School. And it's going to provide a response that generates an entirely new concise description of the grading rubric for Yale law students. It identifies that grades for higher grades are honors pass, low pass, fail, and

that for first-year courses, it's credit, no credit.

But I can actually have a little fun with it as well. I can also ask ChatGPT to do something else like write a lyrical poem. Let me just skip ahead in the video to get to that.

So, I can take that description, as we see, and ask it to write an actual poem. We can give it a minute. So, I type in "now write a short lyric poem," and it can take that description that it generated and actually generate on the fly a brand new poem.

And now, this is a silly example, of course, but there are actually hundreds if not thousands of other real-world applications of people around the world using ChatGPT to engage in conversation, summarize documents, create budgets, create shopping lists, you name it.

And it's not just everyday use cases. OpenAI's ChatGPT, on the other end of the spectrum, can also be used to advance the fields of science in changing ways. So, just to take one example, farmers in India today are currently using ChatGPT to help figure out how to plant their crops. And NASA here at home is using OpenAI models to help search and synthesize their own technical reports.

So, you can see what ChatGPT can do but let's talk about how it works. ChatGPT is powered by a kind of AI called large language models, and OpenAI pioneered this type of large

language model called the GPT model. Now, it's important to remember, why do we call it a model? We call it a model and we use that word because it's not a normal software program or application that's operating in a deterministic manner, as we'd say. What a GPT is, is it's a statistical model that predicts the next word in a sequence by learning patterns in language from vast amounts of texts.

So let's dig deeper on what that means, and we'll start by going through the letters of GPT, starting with G, generative. G is for generative because it's a model, not a database. It generates new content. That's why you sometimes hear this area of technology referred to a generative AI, so let's talk about what that means. Now, you saw a lot here about talking about how training data is compressed and stored in the model and retrieved later, but as I'll explain, that's not correct. That's now how this is working under the hood, and this is a very important concept.

LLMs are not libraries. They're not databases. They don't store information or retrieve documents from some source that they've read. What they do is they create. They generate new texts word by word almost like magic. So, I want to talk a little bit about how that generation actually happens with an example.

So, here's an example of how ChatGPT would generate a response using token prediction by looking at what is the next

likely word in a sequence, and it does this iteratively. At each step, ChatGPT doesn't just look at a single fixed answer, it calculates the range of possible answers. And here's an example of how that might work. If you give an LLM, like ChatGPT and the models that power ChatGPT, a prompt like how do humans learn language, the model will calculate a number of different probabilities for what the first word of that response would be.

Now, there's a number of different ways you could begin this sentence in response to the prompt. You could talk about how they learn from each other or that humans do it by listening to speech or that people learn from example. These are all options that ChatGPT can choose from. But let's assume for this purpose it takes what it believes to be the most likely response and say "through." So it'll take that word and it'll run it back into the model, and it will then iterate over this again. It'll send back the prompt and that first word of that response through to figure out what the next word is, and this process will repeat. And again, there's a number of different ways this sentence may go at this point. It could say through acquiring information, through developing, what have you, but instead we're going to assume it's going to take the most likely outcome of patterns. And this process repeats as it's fed back into the model over and over and over again until you get to a final answer.

Now, if the model had started the sentence in a different way, say with the word "humans," the answer would have come out very differently. It would come out in a different direction, an example of which you see here on the right-hand side of the screen. This is the exact same prompt with two different responses that ChatGPT might be able to generate based on its next word prediction.

And this is a very critical point that your Honor should understood. The answer you get from any given prompt to an LLM is not fixed. So you may get a different output, even when you're using the exact same prompts, and that's because the model is predicting an answer, it's not retrieving one. It is not a database, it is not retrieving some stored document somewhere, it's generating new text.

So, how is it able to do that?

THE COURT: Well, it's generating new text from its databank of a bazillion dillion words.

(Continued on next page)

MR. BAYLEY: That's actually not correct, your Honor. This is an important point because I think -- you get the prisoner in order for us to simplify these issues, we talk about there being words in the model, and I did that here, but it's important to remember what's actually going on in this mathematical model is it's tokens. It's a series of numbers. And so what an LLM is doing when it's looking for a next response, it's actually taking in a sequence of numbers that it sees, and it knows based on the patterns that it's seen before of similar numbers -- it has a series of weights that are baked into it that allow it to be able to develop what is the most likely next token, next number. It's not looking at it as a word. It's not considering it as human language. It's looking at it as a statistical model and figuring out based on the patterns it's seen for this word, or token, I should say, which represents a word, what tokens are most likely to come next. We will get into a little bit of that more as we go through the presentation.

JUDGE STEIN: Why can't I consider a token to be a word or a group of words?

MR. BAYLEY: You can, your Honor. In terms of each token represents a word, that's completely fair. What I am telling you is, in order to operate as a statistical model, it is not retrieving or looking up some particular word as a word as such.

JUDGE STEIN: That I understand. I assume it only does zeros and ones, or something like that.

MR. BAYLEY: Yes.

JUDGE STEIN: Is that essentially it?

MR. BAYLEY: At the base level, yes. But it is all in numerical values that are stored in the model. That's all that's being operated on. That's correct, your Honor.

JUDGE STEIN: For purposes of this litigation, for analysis purposes, do I need to think about that as opposed to considering a token to be a word or a group of words?

MR. BAYLEY: It's important --

JUDGE STEIN: In other words, does it help the analysis from your standpoint for me to think about these as numbers rather than words?

MR. BAYLEY: It's helpful to understand how it works under the hood. Obviously, when it's shown sequences of patterns, those patterns, whether they're represented by words or numbers, they are patterns all the same. But for purposes for my explanation today, the reason I want to orient the Court around numbers is so that you understand what's going on under the hood, because what's going on under the hood is a lot of math. Obviously, you can't do math on words.

THE COURT: Got it.

MR. BAYLEY: That's the purpose of my orientation to tokens.

JUDGE STEIN: Got it. Thank you.

MR. BAYLEY: Let's then take it to the next phase. You heard a little bit about it, but I want to expand and drill down on what this actually means, and that's the training process, and that's the P in GPT, pre-trained. So what is this pre-training process? Pre-training means the model is trained to identify linguistic patterns and that's what happens during pre-training. There are a number of steps to this, so I want to go through them one by one.

One we just talked about, and that's called tokenization. During tokenization, before training can begin, as I explained, you have to take that raw text and you have to transform it into something the model can understand. Computers can only understand ones and zeros, as your Honor knows. So in order to be able to operate on a particular sequence of texts, in order to find patterns, it has to convert that text into numbers. And the way it does that, as I said, is through tokenization. Sometimes tokens -- and the way to think about tokens is it's just a way of chopping up individual words into separate little chunks that can be represented by numbers. And an important point is sometimes those numbers can represent full words, sometimes they can represent parts of words, sometimes they can represent single characters like punctuation or a parentheses. All of it gets converted to numbers before it can be analyzed.

And the other reason I want to focus you on the numbers aspect of this is because of what's going to happen next. So let's take an example of one of these tokens that's developed here on the screen. This is an example of how a sentence -- language develops as the mind learns patterns to be converted into tokens. And if we focus on one of those tokens, "mind," we can then look at what happens next in the process.

So as part of the training process, the first thing you do is you take that individual token, in this case the word "mind," and you expand -- and you deconstruct it. You deconstruct it into a whole bunch of different meanings that can apply to that particular word. So, for example, when you hear the word "mind" in ordinary conversation, that word can end up taking on a bunch of different meanings, depending on the context in which you hear it. It can mean mind the gap. It can mean state of mind in a legal concept. It can mean I don't mind, the friendly verb. It can mean, Mind you. It could mean any sorts of things that could use the word "mind." There is a wide variety of different meanings. This is true of all human language. So in order to capture this wide variety, during the pre-training process the first thing you do is you take this token and you take the meanings of those words and you embed them in a series of other numbers which are sometimes called vectors or embeddings. And it's these vectors and embeddings that are actually inserted into the model in order

to do the training.

And so at the end of this process, by the time you have done this deconstruction, pulled out all the different meanings, the model has a rich contextual understanding of not just one meaning of the word, but all of the different possible meanings of the word "mind" in this case. And again, what's amazing about this process is that these numbers not just reflect the meanings of a word; they also can be used to create associations with other similar-meaning words that are actually different words.

So, for example, if you were to look at the word "brain" instead of "mind" you would see that there's similarity in these numbers, and that's because when the model is looking and trying to understand what a "mind" is, it sees that it has a very close association with the another word called "brain," but only when it's used in particular context where "mind" is intended to be a reference to intelligence or to someone's brain.

JUDGE STEIN: Are you referencing that 26027 is repeated? Is that what you are saying?

MR. BAYLEY: Yes. That's the point -- this is a very simple illustration, but that's the point I am trying to make. When you go into these embeddings, those values that are part of the embeddings, they all represent a wide variety of different meanings and concepts and attributes associated with

that token, and some of those attributes overlap and relate to other tokens. It's a way for the model to learn, Oh, I have seen this word "brain" used in a very similar way to the word "mind." And then its weights, which we will discuss in a moment, are adjusted similarly so it knows when it needs to predict the next word, it can use the word "mind" or "brain" potentially equally, depending on the context.

So you have now broken these things down into embeddings. What's the next step? The next step is you take these vectors, and we have them go through a lot of steps. Now, we don't have time to go through each and every individual layer. It's a very complex process. I will speak a little bit about the overall structure of it, but the core of the issue is, which you heard a little about this morning, is that you show this model billions of tokenized sequences and you ask it again and again, given the sequence of tokens, what's the token that's most likely to come next. And the model --

JUDGE STEIN: That's based on the data set.

MR. BAYLEY: It's based on two things. There's actually two processes, two types of tests that it runs. So what you heard about this morning is the first part of the test where it's shown the data that's in its data set with one word missing, and then based on its answer, it will adjust the weights to figure out whether or not it got that answer right. And if it got it wrong, it will make adjustments to try to make

it more accurate in the future. The end goal of that process, your Honor, is that eventually as part of the training process it will be shown data where it doesn't know the answer, and it will be able to evaluate whether or not it got that answer right. And if it did, it knows that it's not just learning how to predict text it's seen; it's learning how to predict answers based on internalizing concepts and patterns.

So let's take a look at how that actually might work in practice. So here is an illustration of what token training might look like. For example, the model might see this sentence "Patterns emerge from repeated phrases." And it might guess what the next word is. And depending how accurate its guess is, it may need to make adjustments to what are called its weights. It will make those adjustments to those parameters very slightly, and then it will test itself again with another phrase. And based on the accuracy of that prediction, it will make more adjustments of those weights.

JUDGE STEIN: The accuracy of that prediction, in other words, the percentage likelihood that it will match what's in the data set?

MR. BAYLEY: It's close to what your Honor was hearing about this morning called the loss function. So the idea is you have what it knows is the right word, but it's predicting --

JUDGE STEIN: From the data set.

MR. BAYLEY: From the data set. Then it uses that to know, If I am off by this amount, I gotta propagate that back through and make those adjustments to the weights. What's a very important point and what was left out of Plaintiffs' presentation, and I want to make sure your Honor understands, this process does not involve actually storing any of that training data in the data set.

What's happening here is it's taking a set of parameters, I will talk about in a minute how many we are talking about, but it's billions and billions of these individual weights that they start off random, and after this process goes through billions and billions of iterations, these adjustments are made that allow it to make inferences in slightly different ways so that after each iteration, it can get more and more accurate. But what it's doing is, it's not actually encoding any of the content it's seeing; it's encoding the concepts, it's encoding the attributes. It's encoding how people actually learn so that when it sees something similar, it is able to identify it and be able to provide a predicted outcome. It's not remembering, it's not storing the data it's seeing. It's internalizing the concepts and patterns from it. And that's why it's a statistical model that is designed to represent human language, not any specific piece of training data that it sees.

So you do this billions and billions of times over

and, at the end you get a model. Now, as I said, the purpose of this is not to remember -- to try to predict with 100 percent accuracy the word. And, in fact, we will get to this in a little bit. If you do end up predicting the word with 100 percent accuracy each time, that's actually a problem because then it's not being able to actually memorize the concepts; it's only being able to memorize exactly what it's seen before. What you want the model to be able to do and what it's trained to do is to infer concepts that it's never seen before.

I will use another very simplistic example of how this might work. It might involve -- take a simple illustration of the idea of the model being shown a bunch of information about birds, right. And it sees billions of examples of sentences that talk about birds laying eggs. And each time it trains on one of those sentences, it makes adjustments to create an association between the word "bird" and the word "egg" in those billions of parameters and weights that make up the model. And over time, through those adjustments, the model will eventually have, through those combination of weights, an understanding, an inherent understanding that birds lay eggs. And similarly, it may be shown a bunch of texts, a bunch of tokens, I should say, a bunch of patterns that talk about peacocks being birds in various different ways.

Here are a few examples of how that might manifest itself in the training data. It might see that peacocks are

birds, in the pheasant family, they fly short distances like large birds. And through those sentences it again internalizes that association that peacocks are birds. And those associations that birds lay eggs and that peacocks are birds, those are represented as blue and green dials in those weights.

Now, why is this important? This is important because the model can actually go beyond just figuring out those inferences. It can actually come up with new inferences and learn new concepts that it had never actually seen before.

JUDGE STEIN: Let me stop you for a moment. The plaintiffs, as we talked about, were talking about loss percentages and minimizing it and predictive values. You are talking about inferences and almost accessions. You are talking about learning. They were talking about statistical predictions. Is there a difference for purposes of this case in how I should be thinking about it?

MR. BAYLEY: Yes, your Honor, because the output of that prediction is not a stored prediction somewhere. You never actually store the prediction. All you use that prediction for, all you use that training data for is to update these weights that are used to learn and to use to understand these concepts.

So after you do that prediction, you throw that prediction out because all you do is update the weights and then you move on to the next one. You never actually go

through it and keep track of how many predictions you made of a particular sentence or how accurate it was. All that's left at the end of the model is these adjustments to these weights that are designed to represent the concepts.

JUDGE WANG: Isn't the adjustment to the weight essentially learning? I mean, it's adjusting the weight based on the accuracy of that inference, right?

MR. BAYLEY: But it's not learning, you know, that most birds lay eggs carefully in constructed nests. What it's learning is specific concepts like eggs are related to birds. That's what it's learning, and that's the key distinction that I am trying to get across to your Honors is that what's going on in these weights is it's not actually storing individual expressions that it's seeing; it's internalizing concepts. It's internalizing patterns.

JUDGE STEIN: What is any impact on this litigation of that difference?

MR. BAYLEY: Well, your Honor, the impact is there is a theory that Plaintiffs have proffered that all of the training data is stored and encoded and compressed into these models. And the reality is it's just not true as a factual matter, and that will be a factual matter we have to decide here. But what I want to be able to explain to your Honors is what's actually going on. What's actually going on with these weights is it's not actually saying, This is what the training

data said. It's saying, Here is a concept that I have learned from the training.

JUDGE STEIN: I understand what you are saying, and I understand the difference between that and what the plaintiffs are saying.

What is the import of that difference for this case?

MR. BAYLEY: It's the first factor. It shows that what is happening here is transformation. It's taking an individual piece of expression and extracting the concepts. That's fundamental, paradigmatic, transformational use. So this is, to the extent I understand your Honor's point, this is Factor One.

JUDGE STEIN: As I said, it's always helpful to lodge these points in the Fair Use Factor --

MR. BAYLEY: I apologize, your Honor. Factor One. This is 100 percent Factor One.

Getting back to what we were talking about in the slide, the transformational use. One of the transformational uses that is resulting from learning all these patterns is that the model can actually learn new ideas and concepts that it had never actually seen before.

So assume, for example, that this model, even though it had seen lots of information about birds laying eggs and peacocks being birds, it never actually saw directly any example of peacocks laying eggs, but it's able to infer from

common associations that it learned from other things that it had seen and the concepts it extracted from that through its transformation that peacocks lay eggs and it can encode that into the model.

Now, that was a very silly example, but we saw earlier a more realistic example of how this might actually work in practice with the Yale Law School poem. This model may have seen many examples of how the Yale Law School grades work. It certainly saw many examples of poems and how poems work. But I am pretty confident it didn't necessarily see a specific example of a poem about Yale Law School's grading system. But it was able to come up with one anyway. And why was it able to do that? It was able to do that because it had built up enough associations about poems and enough associations about the Yale Law grading system that it could figure out how those associations overlap and create something entirely new, something entirely transformative. That is what's going on with these large language models.

Now, because these pre-trained LLMs, ChatGPT, are capable of inferring new concepts based on what they learn, ChatGPT has an almost limitless number of applications, and that includes solving real-world problems in novel and unexpected ways. This is a fun example I just read about a couple days ago where it was reported that ChatGPT was used to help a group that was lost in some unmapped woods in Nova

Scotia, help find their way back to civilization. And it was able to do this because ChatGPT had inferred through its training data lots of concepts about Nova Scotia, the terrain, landmarks, things that it could use to help those lost ATVers find their way back to civilization. It's just one example of how it can use this transfer learning, as it's called, to have real-world practical applications in a useful way.

JUDGE STEIN: Why isn't that -- my same question. Why isn't that simply understanding how far -- I mean, it having enough data to know how far this point is from that point, in what direction the hikers had to move for how long, that's all just based on straight data, I would -- that's geo location services.

MR. BAYLEY: Right, but this is a completely different way of going about and solving. This is a very transformational way of doing it. Rather than just do what you would do in an ordinary case, which is, Here is two GPS coordinates, give me the results, it's actually able to pull together everything it knows about your surroundings and come up with a very customized description of a path for how you get out. Go down this road. It's going to look like this. You are going to see this sign. And it knows all these things based on other things that it has read about that area and be able to internalize it and know that there's concepts and associations that, Oh, this gravel road is going to take you in

that direction, and, when you get to this particular landmark, that means you are heading in this direction.

It's not using math. It's not using any particular GPS coordinates in that particular way. What it's using is its understanding of learned and associated concepts. And that's kind of what I am trying to explain to your Honor in terms of the transformativeness of this technology. It's doing what may have been done in some ways differently in a new transformative way, using patterns and associations and by learning new concepts.

So I want to then just talk about how to summarize this pre-training process that we have talked about. So to recap, pre-training process describes how the OpenAI's GPT models can transform texts and glean patterns from which it can deconstruct the meaning of human language and infer new concepts. It does this in a number of ways. First, it's got to take that large amount of human language in the form of text data and convert it into those separate tokens, as we discussed that are numbers. It's then going to take those tokens, and it's going to convert them into those vectors, where that vector represents every possible meaning or association that you can have with that individual word. And those vectors are then passed through what are called transformer blocks, which you can think of as simply ways that those numbers are passed through billions of countless layers of processing steps that

result in those adjustments to those weights. And again, there are billions if not trillions of individual weights that are being adjusted every time a single vector is passed through the model. As a result of those billions of iterations, at the end you end up with a set of weights that have been adjusted many billion times, and the result is a model that is capable of predicting with a high degree of accuracy the next word in a particular sequence.

So that covers the pre-training process. And although I have described the pre-training process at the abstract level and made it sound deceptively simple, it's actually quite complicated in terms of how you actually put it into practice. People for decades have been trying to figure this out; machine language, researchers, natural language researchers. This has been a constant problem, and it was only recently figured out by OpenAI how to do it. And it was done by creating this highly complex, multi-layered system called a transformer, and applying it as an architecture as being capable of transforming vast amount of data into the platform foundation we just saw.

So what is a transformer? The transformer refers to the foundational model architecture that defines how large language models process and generate human language. You can think of it like a schematic or set of detailed blueprints. Before you build a thing, you have to have every detail planned in advance carefully. The design of the model architecture

follows the similar principle. Engineers have to create the blueprint when deciding how the model will process information, what kind of computational layers it will use, how those layers are going to interact, how the system will handle specific tasks such as attention, context tracking, optimization, all these things that go into actually adjusting those weights so that they can actually infer concepts and new meaning. And then they write the code and algorithms that bring that structure to life so you can start training.

Now, this transformer architecture is a hugely important part of large language models, and its application is one of OpenAI's key innovations in this space. OpenAI was the first to prove that this architecture could lead to a useful generative language model when it created a model called GPT-1. As part of the transformer architecture, billions of weights, which as noted are essentially just numerical values that are adjusted during the training process, are structured within the GPT model, and they form the core of how the model processes and generates language.

And to give you an idea of scale, OpenAI's GPT-3 model, which was the model it developed before the model it used in ChatGPT, contained 175 billion parameters. That means in GPT-3, every time the model saw a particular piece of training data, each of those weights, each of those 175 billion weights had to make tiny adjustments based on what the model

had learned from just that one piece of text. So how can the GPT model populate and tune so many of those different parameters? How can it tune 170 billion parameters? It needs a lot of data.

So it's important to level-set by what I mean when I say a lot of data. To train an effective LLM, you need a lot of text from a wide variety of sources. That includes text data from the public internet. And the scale of the amount of data you need is almost difficult to fathom. We are talking trillions and trillions of words. Indeed, it requires at least 2 trillion words these days to actually train an effective LLM. And that's actually a gross understatement. Modern LLMs use much more data than even that, but I want to use that just as an example.

To put that in perspective, if you were to start with a news article, which on average represents 500 words, to train a GPT model you need much, much more data, and you need the amount of data on a completely different order of magnitude. Much larger. Not thousands of words, not millions of words, not even billions of words. You need trillions of individual words. So even if you had a thousand books or a massive archives or millions of books and a massive archives of articles, that would barely scratch the surface of what you would need to train an LLM. But that's just the start of it. The other piece is, as I said, you need a wide variety of

diverse text data.

Remember what the goal here is. The goal here is to create a representation of all human language of all human knowledge. That means you have to find other types of language that you won't find in books and news articles. You need examples of colloquial speech. You need examples of folks talking about scientific concepts. You need examples of Gen Z slang, right, if you want to be able to capture the full scope of human knowledge. And that's why you would need a lot of diverse data in order to really be able to make a workable LLM that is actually capable of tuning all those 175 billion parameters.

Now, on this point, your Honor, you heard a lot this morning about the idea of high quality data and what kind of high quality data LLMs need. The concept of high quality and LLMs means something very different than what I think Plaintiffs were suggesting, so I want to make sure your Honor understands. When we are talking about high quality for the purposes of LLMs, we are not talking about the quality of prose, whether something is better written or not.

High quality in the terms of LLM just means, is it something that is actually a representative example of human language? Is it readable? Is it digestible? Is it something that someone would actually say? In that sense it doesn't matter whether a piece of training data was written as an essay

from -- by an eighth-grader or was a Pulitzer Prize winning book. They are all equally important for purposes of capturing all of human language because that is the ultimate goal of an LLM. And to do that, you need a lot of data, and you need a lot of diverse data, and there's no other source of diverse data than really on the public internet, which is why OpenAI uses data from the public internet to train its LLMs.

So let's talk a little bit then about how it does that. How does OpenAI actually access that data? So when you collect the data, it's not enough to just collect the data and download it from a public internet in order to use it for training. And why is that? That's because much of the raw text found on the internet includes content that's not useful for training at all. So take, for example, this Wikipedia page we are looking at. To an ordinary user, you know, as it's rendered in the browser, it appears as a clean, readable piece of text, but if you focus on a particular section of text and you look behind the scenes, the source code that's rendering it, you will see that that page excludes extraneous HTML, metadata, formatting instructions. None of this stuff, none of this stuff is actually usable language that an LLM can actually use. This is what we call in the LLM space low quality data. This is what we are trying to not collect because it is not something that's actually representative of human language.

JUDGE STEIN: Do that again. What is what you are not

trying to collect?

MR. BAYLEY: You see these little HTML tags and these little, like, source code items? Another way to look at it is by highlighting what's actually being used. So if you look at what's actually being collected and used for LLM training, it's only this highlighted portion. All of this other stuff, these tags, these titles, these menu interfaces, these are not things that are actual human readable texts in a linear fashion that someone would actually speak, right. This is text that is useful for rendering a page in an HTML, but it's not representative of human language per se.

And so that's why when OpenAI and other AI companies get this text, the first thing it has to do is run it through a painstaking process to extract the actual language. We call this cleaning. And to clean the data to prepare for training, what it does is we extract the usable text, filter out the irrelevant and low quality noise, which is all those HTML tags that you see here on the page that are not highlighted, and then we apply quality control measures to ensure that the resulting data set is informative, diverse, and is useful as possible.

So that marks the end of the transformer phase, and now you know what a GPT is. You know how a GPT is built. But as you heard, even after pre-training is complete, there's still more work to do. In its raw form, a pre-trained GPT

model simply predicts the next likely words without any built-in sense of right or wrong, usefulness or intent. To create a model with those capabilities, further training and fine-tuning, as it's called, is essential. And after the pre-training phase, therefore, the model goes through what we call a post-training phase, the second phase. And OpenAI was very much a pioneer in this space.

Post-training is where the raw model is shaped into something that is genuinely useful and safe. It requires a tremendous amount of engineering effort to train a model to behave in ways that are useful and safe. So if you take a base model that predicts words and turn it into something like ChatGPT, something that's helpful, it takes an extraordinary amount of effort. It's a complicated process, and as Ms. Maisel said, we don't have enough time to get into the particulars of how it works, but I do want to present one more way of thinking about it, and that's as teaching, right.

So say you are a student learning how to do math. The first thing you do is you give the student a problem and show the student a solution. You are demonstrating to the student how to solve the problem by showing them the answer. We call this supervised fine-tuning in LLM speak. This is where, as part of post-training, it's where you show the model prompts and ideal answers to those prompts. We are demonstrating to the model how to act. And here is an example, a handful of

examples on the screen of how that might work in practice in terms of what prompt might be created and what the ideal response might be shown to the model so that it can adjust its weights accordingly to make sure that its responses in the future are more aligned with those responses.

But let's go back to our student analogy. After you have shown the student how to answer a problem, the next thing you do is give the student a problem and let her answer the problem on her own. And that would give you -- you do that through a process called feedback. And through that feedback, you teach the student how to do things on their own. The same principle applies here. OpenAI created a novel procedure called reinforcement learning through human feedback where we have the model try to answer prompts and we give it feedback that reflects how we want the model to behave so that over many iterations the model gets better and better at a whole range of things. It learns how to be helpful on its own, and it doesn't need to be told the exact answer, and it learns how to operate safely, which is a really challenging thing for a computer, so let's talk about that.

Here we have two prompts: "How to make poison at home" and "Is this mushroom poisonous?"

Now, you and I, we know --

JUDGE STEIN: Go back to 37 for a moment. "How do I make chlorine gas at home?" Presumably the database has

information on how to make chlorine gas. Why is that the ideal response? The ideal response from a safety standpoint, perhaps.

MR. BAYLEY: Exactly.

JUDGE STEIN: But not from a Wikipedia standpoint.

MR. BAYLEY: Exactly, and that's because this isn't Wikipedia. Your Honor, the purpose of fine-tuning a model is to be able to make sure the model is behaving in a way that aligns with human values. That means not teaching -- not assisting people --

JUDGE STEIN: How do you get human values into this? I don't understand.

MR. BAYLEY: Through the fine-tuning process. By telling it which prompts it should respond to and which ones it should refuse. It then internalizes through its weights which responses it is allowed to respond to and which ones it's not. You are basically teaching the model how to be a good chat assistant. And this is an example of a type of ideal prompting response that would teach the model that if someone asks it how to make chlorine gas, it shouldn't do that because it's not safe.

JUDGE STEIN: Well, maybe this isn't the time to unpack that, but it seems that that -- there could be different values, different value systems involved; that is, was Hitler a good man? There are different values in that answer. Most of

us would say no, but presumably, based on what I have been reading, somebody might say yes. How does ChatGPT know what the, quote, "better value" is?

MR. BAYLEY: It's part of this post-training process, and it's part of the rubric that ChatGPT applies. I think we are getting a little far afield. It's an interesting philosophical question, your Honor, I agree. It's a little far afield from -- it's very interesting.

JUDGE STEIN: No, but it's how to build a model.

MR. BAYLEY: It's a constant challenge for us, and it's part of what we are trying to do through this post-training process, is provide a model that is actually aligned with a set of human values, and that is what we are trying to do with the post-training process. I know you have a question of who gets to decide the values, but that's what the training process is for. It's OpenAI is training its model to be safe. It's deciding how to do that.

JUDGE STEIN: If it's not relevant to the lawsuit, just say it's not relevant. Is OpenAI deciding what the right answer to my hypothetical question is? Isn't it based solely on the data set?

MR. BAYLEY: No, your Honor, because again, it's not based on what it's seen necessarily in terms of -- it's not retrieving what it knows about Hitler. What it's retrieving is what concepts it knows, but before it even provides an answer,

there is a separate layer in which it is able to infer, Before I even go through the effort to try to answer the hypothetical, is this a hypothetical worth answering or is answering this hypothetical at risk of spreading disinformation and harm? That's where it will stop, and it won't provide an answer, even if it theoretically could by pulling from what it has learned in the past.

JUDGE STEIN: All right.

MR. BAYLEY: So this is why we think of it as --

JUDGE STEIN: Maybe this isn't relevant to the lawsuit.

MR. BAYLEY: It is somewhat relevant, your Honor, and I will give you an example why. It's relevant because we have a number of safety mitigations that are designed to prevent unwanted behavior, and one of those unwanted behaviors is, for example, reproduction of copyrighted material. So just to skip ahead a little bit.

JUDGE WANG: Can I go back to a less charged example?

MR. BAYLEY: Please.

JUDGE WANG: Before you get to that one, even, for example, the slide with the, you know, "How do I make poison at home? I can't help with that. Yes, the mushroom in the image is poisonous," I am assuming, but you should correct me if I am incorrect, is under -- with, what is, it the RLHF, that there is a human who provides information, and the way that it

manifests in the answers is that the weights of the different concepts get adjusted, right. So the changes, the guardrails, the safety measures. Is there any other way that affects the outputs, other than by adjusting and changing the weights?

MR. BAYLEY: The only thing that affects how the actual probability distribution, the actual, like, Here are the potential responses, that all comes from the weights. So when you are doing the post-training process, just with the pre-training process, you are making adjustments to the weights. What's different about post-training is you are not necessarily teaching it new knowledge per se, but you are rather teaching it how to be useful and safe. You are teaching it how to actually engage in specific tasks which it was designed to do. So for the example with ChatGPT, you are teaching it how to be a useful chatbot, and from OpenAI's perspective, to be a useful chatbot means not to teach people how to make chlorine gas.

JUDGE WANG: I guess going back to that, when you are saying in post-training it's teaching the model how to be useful, is there something else involved, other than changing weights?

MR. BAYLEY: Just changing the weights.

JUDGE WANG: OK.

MR. BAYLEY: Now, there are other mitigations and things that you can also do to improve the safety of the model,

but for the purpose of what we are talking about here, your Honor, for post-training you are just changing the weights. Same as pre-training.

So let's go back to the Slide 39 about the mushrooms. So as I said, we can tell the difference between what is a proper prompt, what is a helpful prompt, and what is a malicious prompt. But these prompts both have very similar words. They both have the word "poison." They both have the -- they both sound very similar to someone like a computer. And it's actually extremely difficult to teach a large language model that has just come out of its pre-training process, based on its pre-training process alone, how to distinguish between the prompt on the left and the prompt on the right. And so that's why OpenAI invented this reinforcement learning through human feedback mechanism, because it allowed through that feedback to inform the model that actually the one on the right is one that you should not be responding to, and the one on the left that you should and, therefore, teaches the model how to be more aligned with human values and behavior. So that's post-training.

So at the end of post-training, you finally have a useful model that can actually do tasks. And what -- the next phase is actually being able to use that model to predict the next word in a useful way, and the model does this through a process called inference. An inference is pretty much what we

saw at the beginning of this tutorial. You give it a prompt and it uses that prompt to predict what is most likely to be the first word in response to that prompt, and then it feeds that back in to generate new texts over and over again until you get to the end.

But as you have heard today, there's a concept that you heard a lot from Plaintiffs called regurgitation, which is a rare phenomenon where a model's output ends up repeating verbatim the end of the sequence that the model had seen during its pre-training phase. I want to talk about what that is and how it happens.

So if I were to say, your Honors, "I pledge allegiance to the," blank, you would know inherently that the next word is "flag." And why would you know that? Because you have heard that phrase countless times. You also would know that because the words "I pledge allegiance to the" is almost always followed by the word "flag." You don't have a lot of other examples of how that phrase might be used. So for this reason -- the very similar concept applies in the case of large language models.

If a model is exposed to the same example over and over again, it will eventually learn to predict that the next word can only be one possible outcome. In the case of the Pledge Of Allegiance here, it will eventually learn if it's only exposed to the same text over and over again, it will only

be able to predict that the next word is "flag." But this is a bug. It's not a feature. It's related to a concept called overfitting, and it reflects the rare circumstances in the model not working as intended. One way to stop this from happening with copyrighted text is to deduplicate the training data, which OpenAI works to do endlessly by removing identical or near similar text from its training data before training ever begins. But, unfortunately, there are some sequences of texts, and this includes, you know, New York Times articles that are republished all over the internet, where things show up in several places, and it can be difficult to mitigate just by deduping alone. So for this reason, OpenAI has developed additional safety mechanisms, and this is what I was getting at earlier.

JUDGE STEIN: Let me stop you. The data set has a certain pattern of words, "I pledge of allegiance to the," and the data set has 99.99 percent of the time the word that follows is "flag." In what way is the model putting in the word "flag" incorrectly?

MR. BAYLEY: It's not about being incorrect, your Honor. What it's about is, you want the model to be able to generalize. It's actually -- when I say "generalize," I mean you want the model to be able to create new texts that it knows based on general concepts that it's learned, and not provide responses based on just a single piece of text that it's seen

before. So I would put your Honor to say it's not about accuracy. What you want the model to be able to do is if it gets shown "I pledge of allegiance," that it would have more variability.

THE COURT: Don't you want the next word in your system to be "flag"?

MR. BAYLEY: Well, it could be -- it could be "flag". It could be "standard." There could be a number of other synonyms for "flag" that could stand in there. But if you have a model that is overfitting on this particular sequence, you are going to lose the variability, and it's only going to be able to predict that one word, "flag." And that's not what you want with a large language model. What you want with a large language model is you want it to be able to have a lot of different possible options, which we saw in the earlier examples in --

JUDGE STEIN: How does it choose amongst them?

MR. BAYLEY: It figures -- it weights all of those probabilities, and then it just -- that sets the possibility. So if it has a 20 percent probability, that means there is a 20 percent chance the model will choose it.

JUDGE STEIN: Wouldn't that give you "flag"?

MR. BAYLEY: It would in this example because this is an example of overfit. You want that percentage to be lower. You want to have more options other than "flag" that are more

possible. That's the idea.

JUDGE STEIN: OK.

MR. BAYLEY: So for this reason, because of this overfitting problem, OpenAI has implemented stringent safeguards to specifically prevent regurgitation of copyrighted content.

JUDGE STEIN: Overfitting is a problem even though it ends up saying "flag"?

MR. BAYLEY: Yes.

JUDGE STEIN: I know you have tried this a moment ago, but do it again. Why is that a problem? When they quote right answers --

MR. BAYLEY: Let me go back to what we were talking about in Plaintiffs' presentation about the loss function, if you remember our conversation. And in that loss function, Plaintiffs kept saying the goal is to make sure it's as accurate as possible, but the difference between the actual answer and the real -- and the predicted answer is basically as low as possible. It's basically zero. If you were to do it 100 times, 100 times you get the same answer. That's wrong.

When you are trying to actually develop an LLM, that is exactly what you don't want to do. You want to make sure it's accurate. You want to lower the number, but there is a lot more nuance predicted because if you are getting all the way to zero, you are going to end up with a model that's not

very useful because all it's going to be able to do is recreate something it's already seen, and that's not what you want. What you want it to be able to do is generalize, come up with new text based on concepts. So when you are looking at that loss function, in reality what you are trying to do is -- you want it to be low, but you want it to be sufficiently -- you don't want -- you want it to be sufficiently far enough from zero that it's not just repeating that text over and over again. You don't want to the weight --

JUDGE STEIN: I understand that's your -- what you are telling me. What is overfit; when it's too low?

MR. BAYLEY: When it's zero. Overfitting is zero. Overfitting is, I am being shown that text so often, my weights are being adjusted so precisely that the prediction is -- the loss function is always going to be zero, and you are going to end up having weights that overfit on that particular piece of data, and are more likely to repeat it in response to a -- as opposed to coming up with something else.

JUDGE STEIN: I would think you don't want the answer to not be "flag."

MR. BAYLEY: Your Honor, the whole idea of these generative AI models is that they allow for all sorts of variety of human expression, right.

JUDGE STEIN: In the OpenAI model, ChatGPT, is there a, quote, "right answer" to what follows the word "I pledge

allegiance to the"? Is there a right answer?

MR. BAYLEY: There is no right answer. There is only a list of probabilities, and we want those probabilities to be as wide as possible to have the most wide variety of possible answers. You want to be able to internalize the concepts of flags and pledging so that it can generate new content, but what we don't want it to do, we don't want it to just repeat what it's seen before over and over again. That's not a useful model. That's not twisting the dials in a way that actually allows it to infer all the concepts that make it be a very functional and useful model.

JUDGE STEIN: I don't mean to overpress, but let me use different words.

MR. BAYLEY: Yes.

JUDGE STEIN: Why would ChatGPT want to give the end user a range of options to "I pledge allegiance to the"?

MR. BAYLEY: Your Honor, in this particular example, it depends on what you are actually asking it to do, right. If the question that is being asked is, "What are the words to The Pledge of Allegiance," obviously, it's going to answer with "flag." So if that is your Honor's question, like, why wouldn't you want it to do that, yes. And that is part of the post-training process. It would learn that that's what the task is. The task is to give an exact answer, and it will then, based on its weights, be adjusted toward that.

What you don't want to do is have a situation -- we are using The Pledge of Allegiance because it's something easy to internalize. It's something we all know ourselves. What you don't want to do is you don't want to get the model in a position where it's just repeating things it's seen over and over again without being asked specifically for that content.

JUDGE STEIN: OK.

MR. BAYLEY: And on the other side, when you get to a point where you are trying to post-train a model into what is actually useful and helpful behavior, you also want to make sure as part of its goal to be aligned with human values that it's not doing things that you don't want it to do.

In this case, OpenAI has implemented stringent safeguards to prevent this regurgitation, especially of copyrighted material. And these protections are designed to stop the model before it even goes -- before it even generates an answer from reproducing excerpts of copyrighted material. Here is an example here on the screen. If you were to type in a prompt to ChatGPT asking it to provide the verbatim text of the opening paragraph from the novel by Jodi Picoult, these safeguards would kick in. The model would know from its training, this is a request to repeat, regurgitate copyrighted content, and it would know it's not supposed to do that, and it would provide a response like you see here on the screen. And why does it show you this? It's because OpenAI has absolutely

no interest in having its models regurgitate the data it sees during training. We want the model to generate new text in new situations. That's why we train the model the way we do to operate the way we do. And this is just one of the safety steps that OpenAI uses to keep its model safe.

We go even further to make sure the models stop regurgitation before it can start. One of the ways we do this is through what I talked about, the deduplication process, to make sure it's not repeating too much. There's also a separate process called -- where we take into account the preferences of content owners, right. One of the ways we do this is through the robots.txt field.

We also do it through something called block lists, which is where the model will know that it's not supposed to train on particular data that it's been told not to. So that includes websites that are owned by The New York Times Plaintiffs because they are on this block list. OpenAI today won't access the website to gather data for an LLM training or search that website to generate responses. And that's why you might have noticed in Plaintiffs' presentation they didn't use specific examples of New York Times content being reproduced.

JUDGE STEIN: That's all in Exhibit J.

MR. BAYLEY: Because of Exhibit J --

JUDGE STEIN: The purported regurgitation.

MR. BAYLEY: Right. Exhibit J is a very specific

example, which I am happy to get into if your Honor is interested. But it was before these were put into place. It's an example of how much work you have to actually do to generate --

JUDGE STEIN: I understand that. We did all that in the motion.

MR. BAYLEY: So with that, I am just going to wrap up real quickly, your Honor, by just going through -- we also have prompts -- other safety mitigations, prompt mitigations, output filters. This is a timeline of the slide. These are the models, if your Honor is interested in knowing which of the ones that actually power ChatGPT.

So with that -- and this is, as you already heard, a discussion of the API. I don't have much to add on that, but to the extent your Honor is interested, I am happy to answer questions about that.

In conclusion, your Honor, I just want to say that, you know, today OpenAI's models, both ChatGPT and API, are used to accelerate scientific discovery, whether it's generating hypothesis, analyzing data sets, they are materially advancing the pace and breadth of scientific research.

So unless your Honor has any questions, I will turn things over to my friends at Microsoft.

JUDGE STEIN: No, that's OK.

Ms. Hurst, approximately how long do you have?

MS. HURST: Your Honor, I have about 25 minutes without questions.

JUDGE STEIN: Go ahead.

MS. HURST: I welcome your questions, your Honor. And, indeed, I am going to try to follow up and address some of them during my presentation.

JUDGE STEIN: Go ahead.

MS. HURST: Your Honor, I want to talk about Microsoft's role in what you have heard, the development of the GPTs, and give a product demonstration that is aimed at demonstrating what you have heard about this morning in ways that will illustrate and provide answers to some of the Court's questions.

I am going to follow up with the discussion of the retrieval augmented generation technology that Ms. Maisel addressed, talk about responsible AI a bit, including some of the mitigations that you have heard about, and then address the issue of old product versus new product that's live in the discovery context for Judge Wang, but also, your Honor, is live in the context of whether the class, the new consolidated class complaint adhered to the Court's admonition to restrict the scope of the case to what was already in the case. So I think that will be relevant to your Honor as well.

So Microsoft's role in enabling OpenAI to develop its revolutionary GPT LLMs is largely comprised of very significant

capital investment and providing the training infrastructure that enables this work to be done. And what does that look like? Well, tens of thousands of specialized chips that you may have heard of called GPUs, very high bandwidth networking that allows for this simultaneous training of these models with all of these weights, and Microsoft has supplied this massive parallel training architecture. And this is what this looks like, your Honor. This photo on the left, this is what one of these data centers looks like from the outside, and the photo in the center is from the inside. It is literally an enormous amount of investment in the technology that is necessary to train and run these models.

So what does Microsoft do once OpenAI supplies it with a GPT? Well, that's just the beginning of the product, Microsoft 365 Copilot, which is at issue in this case. And your Honor, let me make clear, Ms. Maisel was incorrect in her assertion that we have somehow demonstrated the new product. We have demonstrated only the product that was in existence at the time of the complaint and is clearly at issue. We have not demonstrated the new product that was subsequently released, and I will address that in more detail later in my presentation.

So Microsoft takes the GPT, builds a safety system around it, creates a specialized applications infrastructure called an orchestrator, which I will explain in more detail how

that works, and adds the user interface. And with all of those elements, we have what is referred to as Microsoft 365 Copilot. That product is available both -- is available on the Enterprise today for Enterprise users of Microsoft's products. And this is an example of what the user sees, your Honor. The user can choose to interact with additional content on the web or additional content on the work desktop. And that's a part of the M365 Copilot product. So there are some very typical buckets for how people are using these products, your Honor.

JUDGE STEIN: Ms. Hurst, we don't have -- not that I need it, but we don't have a printed copy of what you are doing. Is that right?

MS. HURST: That's correct, your Honor. There are a number of animations here which we will supply to the Court afterwards, but they are better seen, frankly, on the screen, your Honor, than in slides.

So ideation is a very typical example of how people use these products, and it's a really good example to illustrate what Mr. Bayley was describing about the difference between generalization and overfitting. And the reason for that is -- let me give you this example, OK. A couple weeks ago -- it's getting late in the day in this presentation, your Honor, you know, so I hope this will be a little bit amusing.

I was out for fish and chips with my friend at the pub, and it turned out that it was trivia night, so we decided

to play in the trivia competition, and I pulled out my phone, and that was the last time I pulled out my phone during the trivia competition, your Honor, and I asked Copilot what would be a good team name for a two-person trivia team. And I got a list of answers. And, your Honor, this list of answers is a really good illustration of why the training data set is just the starting point for what the GPT model can do. This is a great example of how it generates new ideas. It doesn't just repeat what it saw in the training data. There's no database of trivia team names in that training data set.

Instead, as a result of the training process that Mr. Bayley described, the GPT has learned how to relate different words to one another across a very multi-dimensional space. And so it can go in different directions in that space to generate these new associations. So in this example, Copilot has provided us with some ideas for clever and punny names. I can tell you, your Honor, we chose Smarty Pints that night because we were the punny types. But as you can see, this goes in a lot of different directions. So what's happening here are predictions that are based on generalizations from the data. There's no database there that the prompt is going and saying, OK, give me another example of a trivia name. Instead, there's this multi-dimensional space that all these words live in with this statistical relationships to one another, and when a brand new prompt comes

in, it's able to navigate that space in multiple directions. So let me just pause on some questions the Court was asking of Mr. Bayley.

Your Honor, I am going to say, and I hope Mr. Bayley will forgive me, "I pledge allegiance to the flag" is a bad example of overfitting, and the reason it's a bad example is because there really is only one answer statistically that you are very likely to see in the training data set. A better example of overfitting would be if in all the training data there were five trivia team names, and it only learned those five and it only learned how to spit those five back out, and whenever it was asked for a new trivia team name, it could only give you one of those five examples that it saw in the training data set.

But because that's not what's happening, because there's not a database of the data set, but it has learned the statistical association among words in that word space, the data set is just a jumping-off point for what it can do. It's a jumping-off point for creating that word space, that statistical set of matrices, and then the prompt comes in and the model can go off in all sorts of different directions, depending on how the data, you know -- what the various options are, your Honor.

And let me just emphasize this point about the data set. Even though they are enormous, even though it's, you

know, the amount of words stacked up to the moon and back four times, it is still only a fraction of all of the possible valid word associations that can be generated with human language. And because it's only a fraction, these models don't work unless they are generalized, and you can't get this kind of clever result, you can't get these ideation usage, which is one of the core usages of these models, unless a data set were only a jumping-off point as opposed to some kind of learned fixed pattern. So, your Honor, that's why ideation is a good example of the types of uses that illustrate the ability of these models to generalize and the generalization process that occurs during training.

Let me show you another example. Now, we are going to get into some examples that illustrate the retrieval augmented generation. Your Honor, Microsoft refers to that process as grounding. I think Ms. Maisel also referred to it as grounding. It's because it's using the material as part of the LLM analysis of the tokens. So here is an example of a simple knowledge task with grounding. This is an example where the user is entering a prompt where the user may want verifiable sources of information. It's asking to compare the height of the Empire State Building with the height of the Eiffel Tower. So this is a highly factual inquiry, and one where the user of Copilot may well want to check the sources. And so Copilot provides the user with a variety of sources and links that they

can check on in order to ascertain whether the data that was given with the comparison is correct.

JUDGE STEIN: Do we know what the prompt was in this example?

MS. HURST: Yes, it was, "Please compare the height of the Eiffel Tower with the Empire State Building." So as you can see, your Honor, it retrieved a variety of sources related to that prompt, and it may have retrieved different sources at different times. So we have the Smithsonian, a Wikipedia entry. Other times, I have used this prompt and it's given me the official New York State site for the Empire State Building, and it may retrieve different things at different times. But the point is, this is a task where the user is asking a very factual question and hopes to be able to verify the answer.

So what goes on behind the scenes when this happens is this process of retrieval augmented generation. So as Mr. Bayley described, once the model is completed, it's kind of fixed in place, and whatever factual information that it has gleaned from its training data is then limited to that fixed in point in time. In order to incorporate data that is subsequent to that training process and in order to provide verifiable sources of information, because models do generate, they are not trying to adhere to a specific known set of facts, this additional capability of retrieval augmented generation has become widespread in use with these GPTs and large language

models. And this is the core of the Copilot product that's available both on the Enterprise desktop and in the consumer chatbot.

So how does it work? First of all, you have a specialized type of software and a set of applications referred to as an orchestrator. Let me just get -- we are seeing the same prompt here: "How does the height of the Empire State Building compare to the Eiffel Tower?" That prompt is entered into the orchestrator. Now, this is developed by Microsoft. It's a set of software and applications that interpret the prompt, and made a decision about, is this something where the user wants external data? Is this something where the user intent reflects a desire for verifiable factual information?

And in this case, the orchestrator interprets the prompt as one that is seeking verifiable factual information. And so the model goes to the Bing index, which is the same index that's used for the Bing search engine, and it seeks that verifiable factual information. So first there is a search for information about the Eiffel Tower. Then there is a search for the height of the Empire State Building. And then this information is returned in the form of URLs, in the form of snippet datas, and it's used as grounding material.

Now, what does that mean? What it means is that the data that has been retrieved is then appended to the user's initial prompt. It gets included back within that series of

tokens that are in the prompt, and then the responding model is asked to summarize that information that has been retrieved and provide the answer to the user. So the orchestrator says, We need data for the user. The orchestrator reasoning model performs searches on the search index, it retrieves that data and uses it as grounding material.

JUDGE STEIN: I'm sorry, go back one. Let me see that.

MS. HURST: I am going to have to -- one second, your Honor. My apologies.

JUDGE STEIN: That's all right. You can go forward.

MS. HURST: All right. I apologize, your Honor. We will have to do this a little less animation and more slides next time.

Your Honor, let me give -- I am going to show another example, so maybe we will get it the second time through. Here is an example of a more complex knowledge task with grounding where someone is asking for not just comparing the height of the Eiffel Tower and the Empire State Building. Here, we have an example of, let's say it's a chef and they are going to try to start up a cafe in Manhattan. And, you know, Ms. Maisel referred to this earlier. They are going to try to generate a business plan with the help of Copilot. Now, this is a good example, your Honor, of someone using Copilot to extend their capabilities that they would not otherwise be able to do on

their own. So here, this is our chef. Our chef doesn't necessarily know how to make business plans, so our chef asks Copilot for help. And with just that one query, your Honor, Copilot is able to generate an extended answer with reference to external information to help the chef develop a business plan. So first the answer discusses market trends. And as you can see, there are a variety of links to footnotes that are provided here for the user. It describes the possibility of local competition.

Now, Ms. Maisel helpfully pointed out for me that the Grounded Coffee House had already gone out of business. That's a great example of why somebody doing a business plan using Copilot would use this as a starting-off point and helpful, and would click through on the links and check and see before making important final decisions about how to invest what Copilot helpfully estimates might be a quarter of a million dollars before starting a business. It's a tool to assist someone. And in this case, it provides all of the links, multiple links for all of the data they used in coming up with this complex response of a business plan.

In this case, there were well more than a dozen links that were supplied to our hypothetical chef. You can see here all of the different sources, your Honor. So Sidney orchestrator said, I have got a prompt here that's asking me for a lot of different types of information about how to do a

cafe business plan. I am going to go out to the Bing index and I am going to look for all of the different URLs that may be useful to helping me do that. It selected a number of different sources. It provides more sources at the end, including official government sources and other publications. And then all of that information is summarized and provided back to the user as part of this RAG process, your Honor.

Let me give you an example of the third category that these GPTs are very good at; data analysis. Here, your Honor, we are showing an example using the New York City open contracts data set. As the Court may know, the New York City has published a large data set of the contracts that it has entered into, the money it spends on behalf of the residents of New York City. So we downloaded that data set, and what we found was a spreadsheet that had nearly 50,000 rows and 37 columns; something that would require a great deal of effort to analyze and, frankly, tedium to analyze under ordinary circumstances. And this is a great example of where Copilot can assist the user with data analysis in a much more effective way, reducing the kind of tedium and, frankly, making it possible for people to do things they might not otherwise be able to do.

So here, we load the contract awards data set. This is on the Enterprise desktop, so it's possible, while this is an example of the kind of data that someone might load on their

Enterprise desktop, anybody at work can use whatever type of data they have at work as grounding material with Copilot in order to perform this process. And you will recall, I showed the work in the web tabs at the outset, so this is an idea of what you might use if you were using the work tab.

Let's imagine we are a budget analyst for New York City and we have got this contracts data set and we want to start doing some analysis on this data. So the first thing we have asked Copilot to help us with is, "What's the time period covered by the contracts and the total dollar volume of the contracts awarded?" So you can imagine it takes it awhile to load that data. We have truncated it a little bit.

And let me also say, your Honor, during the process of doing this, we noticed that there were three errors in the data set, so we eliminated those for purposes of our analysis to make it more useful, and also reported them to New York City in case they wanted to correct their data set.

There were 22 --

JUDGE STEIN: How did you know there were errors in the data set?

MS. HURST: That's a great question, your Honor. We noted that the way the data -- the answers that were being supplied by Copilot itself suggested that there was an error in the data. This was actually really interesting. The first few

times we started asking it to analyze the data Copilot suggested there was an error in the data. I started poking into the data set. I discovered that there were three line item entries that were in the trillions of dollars and were clearly erroneous. And rather than try to figure out what the right answer was, I just deleted them for purposes of our analysis, but it was great because Copilot figured out that there was a problem as we were trying to do this analysis of the data.

So the total period of time covered is 22-and-a-half years, and the total dollar volume excluding those three examples is about $276 billion. So now we asked it to please identify the top ten vendors who were awarded contracts; so who is getting, you know, the big dollars from New York City, your Honor. And as you can see here, it has identified the top ten vendors for us: The New York City Economic Development Corporation, The Power Authority, a number of construction companies. And, you know, your Honor, I will say we did sort and look at the spreadsheet to, you know, confirm that this -- looked like an accurate result. So as you can see, there is quite a bit of money being spent on power, on economic development, and on construction.

Now, imagine we are this budget analyst, your Honor, and we are thinking, are we awarding these contracts in a way that is good for the public? Like, are we awarding these by

competitive bid or sole source? What's our procurement methodology here when we are spending $276 billion on behalf of the City of New York? So we ask the next question: "Please provide a table showing the percentage of contracts awarded using each individual selection method."

So, your Honor, as you can see here, Copilot is able to generate a table for us. It shows competitive sealed bids at 29 percent, renewal of contracts of 13 percent. If I was, you know, the budget analyst, I might think, wow, there is a lot going into renewal here. Do we want to look at that? Intergovernmental purchase, that's the Power Authority that we saw. And then it continues to rank all of the rest of the methods that are used for awarding contracts in New York City. And then at each step in the way, your Honor, in this Enterprise product, we can export this in a chart, we can export the business plan into a document and this is the current version of M365 Copilot that is accused in the -- both the class and the news cases.

OK, your Honor. I want to get to this issue of web master controls because it's closely related to the retrieval augmented generation process. And this is important for the Court to understand as part of the complaint of the New York Times and The Daily News and the other news organizations.

So as part of the grounding process, your Honor, as I noted, the orchestrator may seek information from the Bing

index, which is the very same version of the index that is used for the Bing search engine. And here is an example of visiting the Coursera site. Let's say the user has prompted asking for information about technical training courses, and you see that Coursera has come up as an answer, and the information about some classes are available, but, oh, we have hit a paywall. We have hit a paywall.

Now, in retrieval augmented generation, as it's implemented in Copilot, your Honor, the process includes respect for these kinds of limitations. There is a paywall. The code reflects that there is a paywall. It says, You have got to be able to sign in to get any further access to this data, and so we are not going to allow you to go behind the paywall to get that data. But paywalls are not the only way in which these limitations are implemented, your Honor. They are also implemented through meta tags. And you can see here on the screen two meta tags that have been published by Microsoft -- they were published before this lawsuit was ever filed by The New York Times -- called No Cash and No Archive. Let me explain to you how these work.

When The New York Times publishes its digital edition, it has all that text coding and HTML that you saw when Mr. Bayley was showing you the difference between quality data and chaff. Well, in the chaff, your Honor, you can implement this instruction, No Archive or No Cash, and you can choose to

exclude your data from use in the retrieval augmented generation process entirely in Copilot by implementing these tags. So at anytime prior to and during the pendency of these lawsuits, if the news organizations did not want their articles or snippets of their articles or even URLs to their articles included in grounding data, all they had to do was put these tags when they put it up on the internet. That's all it took.

And, your Honor, I don't want to add to the discussion about Exhibit J very much here today, but -- because we said -- we heard from Plaintiffs that they would never talk about it again, but nevertheless, this is why we are so troubled by what's in Exhibit J when it comes to Copilot, because we know we have implemented these controls. And I am going to talk in a minute about the other safety measures. And nevertheless, they are showing things that we cannot reproduce, and it causes us great concern about the inclusion of those examples in the case.

JUDGE WANG: Ms. Hurst, I had a question. When you talk about -- and this might have to go back to the slide with the bigger and bigger squares that starts with a small GPT square.

MS. HURST: Yes, your Honor.

JUDGE WANG: When you say that the process -- the orchestrator may seek info from the Bing index, but then the process includes respect for the paywall, can you tell me which

box that falls into? And then similarly, I think what I am hearing from your discussion about meta tags is that the meta tags need to be put on by the content owner.

MS. HURST: That's correct, your Honor.

JUDGE WANG: Right. So those meta tags would -- where would those be respected, I guess?

MS. HURST: Yes. Happy to explain that, your Honor. So as I noted and as you can see here, there is the Bing index, and in the Bing index, that's basically a cached copy of the internet, right. So when you operate a search on Bing, or any other search engine, it's not going out on the internet to look for what you want for the very first time. It's got a cache copy of the internet and it's looking at that cache copy and then giving you the links that it was last aware of that are most relevant to that information. And the same thing is happening here as part of the Copilot process.

So during the process of retrieval of those links to be included as grounding material, your Honor, there is a check for one of these tags, and if the tag is present, it's excluded and it's never even provided to the model to be used as grounding material. So it's excluded at that first check step where the model looks to see, is there one of these meta tags? If so, the orchestrator doesn't even return that data, and it can't be used as grounding material and providing the output to the user.

Now, your Honor, I am mindful of Judge Stein's queries repeatedly, Why does this matter, right? It matters in two ways to the defenses in this case. First, it matters as part of the third factor because it demonstrates that Copilot is not using any more than is necessary to answer a user's question. But it also matters for a separate defense, which is the implied license defense, your Honors. And that implied license defense will have a significant role in this case with respect to the retrieval augmented generation technology.

It's long been the law that search engines can make copies and provide snippets and URLs even though those snippets might substitute for a click-through -- that's what the existing search engine cases are about -- and that web masters can offer these kinds of tags to allow anybody to exclude themselves from that process. And that implied license defense related to that has long been established in the law as well. So it's relevant both to fair use and to this other defense. And so at anytime the Plaintiffs news organizations could have taken advantage of these tags in their digital publications. They all have accounts with Microsoft where they use web master controls. They are aware of these controls, and they can use them.

And by the way, your Honor, just to be clear, Microsoft is not saying, We won't put you in our search index unless you let us use your stuff for AI. These are dials where

people can turn up or turn down their willingness to participate. They can say, I will let you have my URLs, but I won't let you have anything else; I will let you have a URL and a snippet, but I don't want you to have the whole document; or they can leave off the tag and just have it be freely available as grounding material, knowing that people will click on those links and they will look for more information.

So, your Honor, Mr. Bayley talked about some built-in mechanisms to try to make the GPTs safer in practice. Copilot also includes a lot of safety layers built on top of that. It includes a front-end check of the prompt to make sure that it's an appropriate prompt. At that point in the process, it can stop, short-cut and say, I don't view this as an appropriate prompt and I am not going to respond to it. That might include, you know, things related to self-harm or, you know, weapons or violence or other kinds of material.

It also checks both on the front end and on the back end, once a response is generated, for a comparison to third-party content. So, for example, your Honor, on the front end, if the prompt says, Please give me the verbatim text of such and such an article for the Orange County Register, that's supposed to be an immediate disrupt and not occur. And the same thing is supposed to happen on the back end with respect to the filter check, where it looks at what the output is and compares it to the grounding material and says, This should not

be a verbatim copy, or it should not quote extensively from the grounding material. This is why, your Honor, we are so concerned about Exhibit J. And I don't want to belabor the point any further, but we had -- in order for what the plaintiff showed from Copilot today to have occurred, they have to have failed to apply the tags, and then somehow gotten around every one of these mitigations. And that is why we have been so concerned about it.

OK. Judge Wang, I am getting to that part that is why is the new Copilot different. So as I mentioned, the original Copilot, which is still in operation and which we have been looking at here today, which is called now Microsoft 365 or M365 Copilot, is available on the Enterprise desktop. It was first called Bing Chat, and it was launched as a consumer product as an advanced search product that brings this GPT technology available to the masses, all the kinds of things that we saw today, in an advanced search capability for free. And that was Bing Chat. And it was rebranded as Copilot in November 2023, and it was also launched on the Enterprise during this period of time.

Then we have, just for purposes of illustration, where The New York Times complaint is filed. At that point, Bing Chat, the version that was rebranded to Copilot, subsequently adopts GPT-4. And we all agreed that GPT-4 and that version of Copilot was in the case. So there was no dispute about that.

What there is a dispute about, your Honor, is New Consumer Copilot. New Consumer Copilot was released in October 2024. That's the free chatbot. It first used OpenAI's version, GPT-4o, which Microsoft never agreed was in the case as to it; GPT-4o or New Consumer Copilot. Now, why does this matter? Well, because it was a product that is completely new and was coded after the lawsuit was filed. And, of course, the old -- even the old product has been continually developed during this period of time and is still in operation. But what really matters with respect to New Copilot is that it's completely different.

Judge Wang, you asked the question yesterday, Well, if it's got the same name, you know, maybe -- and it's basically the same, maybe I should treat it as the same. I promise you we had this slide even before you asked that question, your Honor. And this is the answer, right. The Shelby GT350 from 1966 is called a Mustang, and the 2024 Mach-E is also called a Mustang, but they are totally different products. They are different in design and they are different under the hood also.

And here are the ways in which New Consumer Copilot are different under the hood. So it's an entirely new infrastructure and system architecture. It's built on a different platform, and it's got a different logging system and a completely different collection of logs. So Judge Wang, when we have been talking about that 15 petabytes of output data,

that is not this new system. It's only the old system, your Honor. And it's got a new code base. Well, why does that matter? Plaintiffs have already spent 50-person days examining the code base. That Sidney orchestrator that I showed your Honors, that code about how does it decide when to go to get grounding material, does it check, all of those things, they have already spent 50-person days, and we would be starting all over on the new version were the New Consumer Copilot to be in the case, your Honor. Those are noncustodial data sources, your Honor.

With respect to custodial data sources, there are -- this code was written from scratch by about 300 people starting in May of 2024. Only two of them overlap with Bing Chat, so it's a different group -- that's not even counting, like, the product marketing people. Your Honor, Microsoft acquired a whole new team from a company that was formerly called Inflection in the beginning of 2024, and it was that team that built this new product. And there's very little overlap between the existing number of custodians and that group of people. So that's, your Honor, to address Judge Wang's question, but I'm sorry to say, Judge Stein, you may also be seeing some issues related to this in connection with the scope of the class complaint.

So let me sum up here and then also, of course, be happy to answer any questions the Court has.

Mr. Bayley explained how GPT for OpenAI refers to generative pre-trained transformer. Microsoft also uses the acronym GPT to refer to this technology in another way, which is a General Purpose Technology. And I showed you some very mundane examples today of trivia team names, comparing the heights of things, developing business plans, data analysis. But the capabilities that are reflected in those examples have allowed people using Microsoft Copilot at Johns Hopkins to detect and prevent disease, other scientists to speed up materials in scientific discoveries, for Khan Academy to create two tailored tutoring plans, math tutoring plans for students, for the UN to adopt a new worldwide system of disaster warnings. We take for granted here in the United States we know when a tornado, a hurricane is coming. In a lot of parts of the world they don't have that. Copilot now helps supply that.

And what's the significance of this being a general purpose technology? Well, at the outset, your Honor, I told you that Microsoft had invested billions of dollars and provided a massive amount of capital infrastructure to OpenAI. It did not do that because it thought it was funding the development of a regurgitation machine. General purpose technologies have been the spur for human development through repeated advancement. In the First Industrial Revolution, it was iron working; in the Second, electricity and mass

production; in the Third, computerization and mass digitization.

Your Honor, large language models are general purpose technologies that Microsoft believes have the capability to fuel that fourth revolution, and it is invested accordingly, and we believe that this technology is a fair use of copyright and is appropriate with respect to the other legal challenges that have been offered by the plaintiffs.

And I would be happy to answer any other questions from the Court about training, retrieval augmented generation, or any other subject.

JUDGE STEIN: I am fine. Thank you. We are fine. Thank you very much.

I have quick questions of the plaintiffs in terms of not simply regurgitation. What about the example that was given of the two-person trivia team, with all of those new answers, clever and punny, and so forth? So that certainly substantially differs from regurgitation. That's one. Quick answer, please.

And secondly, why doesn't The New York Times, unless it does, put the meta tags to say, Don't allow the copyrighted material to be used in your machine?

MR. NELSON: Thank you, your Honor. Justin Nelson.

Let me answer the first question on regurgitation and the pun aspect to it. Remember on the loss function, on the

base model, it is to compress to have an accurate representation as close as possible to zero. And it is true that LLMs generally do not reach a loss of zero during training, but companies training the LLMs are constantly optimizing their models to get as close to zero as possible. You recall that graph, if you go back to Slide 34, it is at -- it is not at zero, it's just quite low, and it's done by the scale of the data and the scale of -- Mr. Smyser, can you --

JUDGE STEIN: I have it. Go ahead.

MR. NELSON: It's a scale of the data and it's, of course, the power of the compute. And so they are trying to get as close to zero as possible, even if they are not going to approach a true lossless compression.

On the pun part of it, that is on the output side. It is still based upon -- Ms. Maisel described sort of the temperature setting, and so they are using the base model that's still a compression. That model is a compression. And then on the output side, when you ask it something, it uses the user's input to then suggest names based upon the expressive content that is embedded in the model.

I hope that helps. I am happy to go into more detail about that.

JUDGE STEIN: No, just redo it because it doesn't strike me as simply regurgitating with a near zero loss factor to have all of those -- I forgot what they were, but, you know,

clever puns that involved trivia and brains, and so forth.

MR. NELSON: Well, your Honor --

JUDGE STEIN: You can talk to each other if you want.

MR. NELSON: I am getting a note, but before I say, I want to confirm its accuracy.

Just on a quick Google search I am informed that, for example, Quiztopher Nolan --

JUDGE STEIN: Just a minute.

Go ahead.

MR. NELSON: It's going to come up on the screen now, but actually it's not a made-up name. According to what we are seeing right now it's from the web. And so "Small batch trivia at dive bar. Geeks who drink." Look at 67: "Quiztopher Nolan." Again, this is just off of literally in the courtroom right now, having literally, as we are trying to answer your Honor's question. And so the point is, they are using the underlying training that exists.

JUDGE STEIN: All right. I got it.

What about the meta tags?

MS. MAISEL: Your Honor, Jennifer Maisel again. I can't confirm for every single article, but I do know for some of the articles cited in The New York Times complaint, the meta tags that Ms. Hurst said would prevent the display of New York Times articles in the output were in place at the time of the complaint. They were in place at the time that output was

generated, so they didn't work. Those meta tags were in place and The New York Times content, the web pages that had those meta tags, were still displayed.

And I have two other points about the meta tags, your Honor, just very brief.

JUDGE STEIN: OK.

MS. MAISEL: Ms. Hurst did say that you can implement the meta tags and that will not exclude your website content from search. What she didn't say is that those two tags, the No Cash and the No Archive tag that she displayed, presents a very difficult choice for publishers. One of those tags allows publishers to show up in the generative AI search answers, but a condition of that is, you have to make that content available to Microsoft for AI training.

Alternatively, you can opt out entirely from both AI training and display in those generative AI answers, but you won't be shown in search results then. These Bing search results, these generative AI search results from Copilot, they are shown at the top of the screen. All the links follow and then traditional search results. So it's a difficult choice. If most of your traffic is driven by search as a publisher, you are faced with an impossible choice: Either give up your content for training by Microsoft and for all sorts of other purposes or don't show up in search results, and have people drive -- and have those search results drive traffic to your

site.

Just one other point I want to make. These meta tags existed long before Microsoft announced these controls. These controls have not been widely adopted in the industry. They are very new. So to suggest that they are standard or well-understood behavior, I think --

JUDGE STEIN: Well, they are available was the point.

MS. MAISEL: Yes, but they are not -- there is a very difficult choice to make in terms of --

JUDGE STEIN: OK. Thank you.

All right. I thank everybody. Here is what I would like to do. We will have a very brief argument on OpenAI's appeal from the preservation order, and then I will give you some time for lunch, and then people should go to Magistrate Judge Wang's courtroom for the, I guess it's the discussion of the preservation order. All right.

JUDGE WANG: For anybody who is attending the settlement discussion, it is not open to the public. The lawyers are permitted to bring food and drinks, so I don't want anybody to miss lunch. I don't want anybody to be hangry when we are having a conversation.

JUDGE STEIN: Because we are so concerned about you, I will give you time for lunch, but no food in this courtroom.

Thank you, Magistrate Judge Wang.

MR. VAN NEST: Should we pause for a minute to give

folks a chance to leave?

JUDGE STEIN: Yes, anybody can leave who wants to.

MR. NELSON: During the transition, one quick housekeeping matter.

JUDGE STEIN: Let's wait.

(Pause)

MR. NELSON: One quick housekeeping matter. Would it be acceptable or would the Court like us to file as a notice or as a docket the slides that were used in the presentation today?

JUDGE STEIN: I have them. I have no feeling one way or the other. I think it would be generally helpful for the public to have access to them.

MR. NELSON: Thank you, your Honor.

JUDGE STEIN: All right.

MR. VAN NEST: That's fine with us, your Honor, as well.

THE COURT: Go ahead. So the parties will docket those.

All right. We are talking about the May 13 preservation order. You lodged an appeal. You realize the standard under Rule 72 is clearly erroneous or contrary to law. Pretty high standard. You know that there are discussions that have either been going on or are about to go on in regard to trying to come up with some resolution of that order, but I am

prepared to hear the appeal right now. Go ahead, sir.

MR. VAN NEST: Your Honor, good afternoon. Bob Van Nest again from Keker, Van Nest & Peters for OpenAI. Thank you very much for giving us an opportunity to be heard on this important issue. We have tried to be very selective. I know Judge Wang is working very hard, your Honor, as well.

Judge Wang is continuing to manage this issue, including today with a settlement conference discussion. So our objection for your Honor's review today is limited to the order requiring OpenAI to continue to preserve output logs for users who have expressly asked us to delete those. There are two bases for our objection, your Honor, as to why Judge Wang's order is erroneous in this regard. First of all --

JUDGE STEIN: But isn't it a permissible inference that end users who may believe they are violating the copyright laws would want the -- not to preserve evidence of their attempt at vicarious copyright infringement -- actually, copyright infringement, so that -- although I don't know what the answer is -- the percentage of copyrighted material generated in response to prompts by people who have checked, I want privacy, I want these deleted, is greater than in the general population? Isn't that enough to permit the preservation order, as limited as it may be in time, to go forward? That's loaded, but go ahead.

MR. VAN NEST: Of course, your Honor. I don't think

so. I think, given the impact this is having on OpenAI and its users, I think you need more than inference to support that kind of a broad order. Let me explain why. So far no one has come up with any support, any evidence that that inference is actually true. And this is an unusual situation.

JUDGE STEIN: There can't be -- isn't the evidence there? I mean, it hasn't been gone through.

MR. VAN NEST: We have produced hundreds of thousands of pages of documents, e-mails, a backup, so on. There is no evidence of any of that in there. And --

JUDGE STEIN: Where the privacy restrictions were put on?

MR. VAN NEST: No. In other words -- but, your Honor, this is unusual because the default for ChatGPT is we do retain the data. The default is we do retain the data unless, unless the user says, either by activating a temporary chat or deleting a chat, that they want their privacy. And we commit to those users that we will do that. We have made a commitment.

JUDGE STEIN: That I understand, but permit me the inference. You can push against it. Permit me the inference that someone who is concerned that they are violating the Copyright Act might, indeed, be more likely to effectuate that delete than --

(Continued on next page)

MR. VAN NEST: There are so many --

THE COURT: -- than the user in general. And in terms of the affidavit you have, there's really no specificity on the cost of what that costs you to do that.

MR. VAN NEST: Let me put the cost aside. I'll address that separately. But as your Honor knows and as, I think, Judge Wang recognized, there are many reasons why users elect to preserve their privacy. They are chatting about personal financial data, deeply personal family information, confidential business plans, embarrassing information of all kinds. There are many reasons why users choose this privacy.

And everyone acknowledges that the percentage of this information that's likely to be relevant is vanishingly small. Even the plaintiffs acknowledge that we're talking about six-one-hundredths of a percent. Now, we've already retained billions and billions of chats.

THE COURT: Wait. Six-one-hundredths of a percentage?

MR. VAN NEST: Six-one-thousandths of --

THE COURT: Is what?

MR. VAN NEST: The likelihood that there's any even relevant data in ether the set we have retained, which is billions of chats, or the set where users have requested privacy and we haven't retained it. They're acknowledging --

THE COURT: All right. I'll want the other side to respond that.

Go ahead.

MR. VAN NEST: Yeah. They've acknowledged that the likely percentage of anything relevant is vanishingly small, which is why I say in this -- usually, the default is you get rid of the data. Usually we're here because someone has deleted data, that's the default. No, our default is we retain the chats and we commit to our users that they can make a privacy election. And as I said --

THE COURT: I understand. But what you're saying in response to my inference is really very little chance of that inference being effectuated. That's what you're saying.

MR. VAN NEST: I'm saying two things, your Honor. I'm saying there's very little chance that there's anything relevant, and I'm also saying there's simply no support for that inference in the evidence at all because there's --

THE COURT: So, how about logic? If you think you're doing something wrong, you're going to want that to be deleted.

MR. VAN NEST: I doubt very much whether, in this instance -- the wrong thing you're talking about is using ChatGPT to get a copy of something; right? That's what --

THE COURT: Yeah.

MR. VAN NEST: The public at large probably doesn't regard that as very wrong; right? That is a copyright construct that we're litigating here in court, but it's not like somebody's committing larceny or bank fraud or wire fraud;

this is something that folks out there in the world maya or may not do, but whether that rises to, hey, I'm going to conceal this, there's simply --

THE COURT: The plaintiffs are saying that you're shooting them through the heart. I mean, you know --

MR. VAN NEST: Well, it would be one thing --

THE COURT: -- it's how it's viewed.

MR. VAN NEST: They're saying and they're conceding that the amount of data in either of these two stacks that is potentially even relevant is very, very small, and that's logical because we're talking about billions and billions of chats. We have saved billions and billions of chats and we are working closely with Judge Wang to figure out how to sample those. Right? That has to be sampled. We can't go through and -- that's not something that can be searched.

And I would say that the problem with the order is it has the burdens wrong. Right now we, ChatGPT, OpenAI, are being punished. Our users out there are being told that they're on a lit hold.

THE COURT: Wait. The burden on this motion is on you.

MR. VAN NEST: It is. But the burden of establishing the order was on them; right? And the way the order was entered -- and this order was entered without a response from us initially and then we had to move to reconsider. So

Judge Wang acted as soon as they got a misleading pleading from the plaintiffs and then we had to move to reconsider.

But what the burden on us is -- let's put the cost aside. That could be very, very expensive too, but I accept the fact that it may be temporary. The current imminent burden on us is our users are being told they can no longer preserve their privacy. We made a commitment to them that we would do that, we have been honoring that. Now that we have the order, we are preserving --

THE COURT: If I remember correctly, the section that you have says unless ordered by the Court or unless for legal reasons.

MR. VAN NEST: Sure. I'm not saying --

THE COURT: An order of the Court is a legal reason.

MR. VAN NEST: It is. It is. However, it is the equivalent to putting millions and millions of users on a lit hold. And people who have been using ChatGPT for their personal information, their health information, that's the real --

THE COURT: I understand that. I'm cognizant --

MR. VAN NEST: Now, there's a cost issue, too, but I appreciate that this order is intended to be short-term, and Judge Wang has said it's not going to go on in perpetuity, and we hope it'll be resolved quickly. But my point is this issue is urgent because out there in the community of chat users,

they are aware that suddenly all of their chats are now being preserved, and that has a huge impact potentially on our business; right? People are not going to use ChatGPT.

THE COURT: Well, what evidence do you have for that? By the way, I do not mean to belittle those privacy concerns in the least. They are clearly important to people, and you have given them the option of effectuating their privacy concerns. This is an issue of how long it's going beyond, the fact that it's being negotiated now and this afternoon, and the fact that it's temporary.

MR. VAN NEST: And I accept that.

THE COURT: And presumably, there are ways to both anonymize anything that's retained and undoubtedly there will be sampling, as you say. I think all of those mitigate the privacy concerns.

MR. VAN NEST: They may mitigate it, your Honor. But I would flip it around and say, if we're only talking about a week or two, there is no harm in allowing OpenAI to honor its commitment to its users. That amount of data in the next couple of weeks is not going to make any difference to this litigation whatsoever. And if it's true, and I think it is, that Judge Wang and the parties are working hard to get this done, then okay. Another couple of weeks is not going to matter to the litigation. It does matter, every day matters, to OpenAI because, again, our users are out there and we aren't

honoring our promise to them.

THE COURT: You are honoring your promises because you've said we're going to hold onto it if we have to. If we're directed to.

MR. VAN NEST: Understood.

THE COURT: Okay.

MR. VAN NEST: But for them, it's like being on a lit hold and they're not even parties to this. They're innocent third parties out there using ChatGPT and they're all being told you're on a lit hold now. And there's no --

THE COURT: You're on a what? A lit hold?

MR. VAN NEST: Litigation hold. Yeah.

THE COURT: Yeah.

MR. VAN NEST: There is no termination date now on this order. That's clear and that's a concern.

THE COURT: Trying to get that resolved quickly.

MR. VAN NEST: Well, I hope so, and we're trying to do this. The plaintiffs keep throwing up roadblocks to our sampling proposals and that's part of what has extended this: Roadblock after roadblock after roadblock. I don't think they really want the answer. I think they know the answer is going to be no, there's nothing out there like this.

But my point is right now it's open-ended. There is no end date. We're working hard to get it done. But if it's true that it's going to be short-lived, there's nothing wrong

with letting us go forward, reversing that part of the order only, just the hold --

THE COURT: That data will be deleted by definition.

MR. VAN NEST: Right.

THE COURT: That's the cost.

MR. VAN NEST: Yeah. For the next couple of weeks or whatever it takes. A very short time. What I'm saying, your Honor, is that the cost of that to this litigation is nil, and that's all.

THE COURT: I understand.

MR. VAN NEST: Yeah. Thank you.

Quick response. I don't think you need a weatherman but let me hear from you.

MS. MAISEL: I'll be brief, your Honor.

Jennifer Maisel from Rothwell Figg on behalf of the News plaintiffs.

So, to directly answer your Honor's question about the six-one-thousandths percent rate, I direct your attention to page 19 of News plaintiffs' response. This is docket 212.

THE COURT: Sure. Just a moment.

Yes, ma'am. Go ahead.

MS. MAISEL: There's no concession that the six-one-thousandths percent rate is accurate for this case. As your Honor mentioned, we do not have the data. We don't know what the prevalence rate is, and OpenAI has not told us what

the prevalence rate. And when I say prevalence rate, I mean the rate at which news is used as part of the ChatGPT or the API products. Instead, this number came from a different litigation, from a different defendant, that defendant was Anthropic, another generative AI company. It is the prevalence rate that Anthropic offered to the music publishers in that case for the prevalence of users asking for song lyrics in Anthropic prompts and outputs.

And as a baseline for purposes of coming up with this sampling proposal, News plaintiffs suggested, why don't we start with that because we don't know what that prevalence rate is. And earlier on the same page, page 19, we have OpenAI's head of product, Nick Turley, saying in the context of another case, the antitrust case against Google in remedies hearing, he says that local news is a pretty common query in ChatGPT.

So, I don't know how "pretty common" correlates to six-one-thousandths of a percent, but by no means have we conceded that that is the actual prevalence rate, it's just a baseline for trying to come up with a sampling methodology.

And just to briefly respond to some of the other points. Your Honor was correct about anonymization, and I do want to make clear yet again on the record that we are not seeking specific users. We don't want to know what someone put into ChatGPT to get an answer about buying birthday presents for their toddler or other personal --

THE COURT: Well, no. The reference was deeply personal information which undoubtedly exists.

MS. MAISEL: And we're not after that. We're after the copyrighted news articles and the issues present in this case. We proposed search terms, we've been going back and forth to try and target that type of content in the output logs, but we're not interested in that other kind of data, nor do we want it. And we consent to appropriate safeguards to anonymize and deidentify that information and otherwise filter it out.

So, there's no question that we're not willing to work with OpenAI and want to work with OpenAI to address those privacy concerns, so there's no issue.

THE COURT: All right. Thank you.

Now, I am going to affirm the order. The News plaintiffs sought and order directing OpenAI to preserve all output log data. On May 13 the magistrate judge directed OpenAI to preserve and segregate all output log data that would otherwise be deleted. We've talked about that today. The burden on this motion is on OpenAI.

I think it's a permissible inference to believe that a certain percentage of users, whatever that may be -- we don't know, we may find out -- would have an interest in deleting the output log data because they believe they may be engaging in copyright infringement. I don't know what that number is, but

I think that data should be retained. That is, I agree with the magistrate judge's order. It's going to be temporary. This motion, in a way, is a little premature because the parties are attempting to resolve that issue. Ultimately, any information will be sampled and will be anonymized.

And I think most importantly, as a nondispositive order, the order is to be reversed only if it's clearly erroneous or contrary to the law. That's Federal Rule of Civil Procedure 72(a) and 28 U.S.C. 636(b)(1)(A). So, for all those reasons, I'm denying the objections and I'm affirming the magistrate judge's May 13 order.

I thank everyone for being here today. I kept you a long time, it's 2:00 now. Do you want to take a half hour before you go to Magistrate Judge Wang? I'll let her know that, that you'll be there at 2:30. I don't think there's any issue about bringing food in because I think the cafeteria closes at 2:00, but I'm not sure. Something tells me if I go down, I'll get my lunch, and perhaps the parties will as well.

I want to leave you, first of all, with thanking you and saying just a few words to you, and those few words are August 14, 2026. That is the day in the scheduling order that the parties are to file their motions for summary judgment. As I've said, I think that's where this is going. I really don't know what the answer is going to be, but I think the summary judgment motions will be very important here.

I have no current intention of changing that date. I'd like to say I have no future intention of doing it but I don't have the ability to do that. August 14, 2026. I leave you with that.

Thank you all.

(Adjourned)